ART AND SOCIETY

ISSN 2709-9830 JUN. 2026 VOL. 5, NO. 2

CONTENTS

Paradigm Academic Press
Art and Society
ISSN 2709-9830
JUN. 2026 VOL.5, NO.2

Navigating the Ontological Blur: A Stylistic Inquest into Poe's *A Dream Within a Dream*

Mark Kevin Astrero[1,2]

[1] PhD Student in Rhetoric and Linguistics, St. Paul University Philippines, Tuguegarao City, Philippines

[2] Faculty Member, Northeastern College, Philippines

Correspondence: Mark Kevin Astrero, PhD Student in Rhetoric and Linguistics, St. Paul University Philippines, Tuguegarao City, Philippines; Faculty Member, Northeastern College, Philippines.

doi:10.63593/AS.2709-9830.2026.06.001

Abstract

This study presents a comprehensive stylistic analysis of Poe's 1849 poem *A Dream Within a Dream* that surpasses traditional approaches to Poe analysis, which focus primarily on biographies and themes. The gap in the existing body of Poe's work centers on how isolated linguistic elements of style have not been synthesized for a complete analysis of Poe's work as a body of literature. To fill this void in the current literature on Poe, this study employs a strict three-tier analytical model that integrates Sense, Sound, and Structure into one cohesive analysis of the poem. At the level of sense, linguistic style elements, including figurative language, denotation/connotation, and sensory imagery, are examined to illustrate how the poem translates the abstract quality of transience into the tangible reality of the physical world. At the level of sound, prosodic features are examined to demonstrate how the poem's sounds recreate for the reader the continuous ebb and flow of time while simultaneously inducing feelings of profound existential despair. At the structural level, irregular line lengths, variable stanza lengths, and extensive enjambment are examined to show how the poem's structural characteristics mirror the instability of human perception. Overall, the integration of these three disparate linguistic tiers demonstrates how Poe's intentional use of form creates an active representation that raises the question of an ontological blurring of the lines between reality and illusion. Finally, this analysis contextualizes Poe's poetic language within an applied linguistics framework, demonstrating the structural relationship between aesthetic inquiry and the integrated form of an aesthetic object.

Keywords: stylistics, Edgar Allan Poe, poetic rhetoric, prosody

1. Introduction

In this paper, the significant relationship between the form of literature and the meaning of language is explored by applying a detailed stylistic analysis to Edgar Allan Poe's lyric poem, *A Dream Within a Dream*, written in 1849. The chapter uses an interdisciplinary approach, grounded in applied linguistics, to introduce the reader to the study's research material and to show how poetically structured language constitutes an interrelated system that produces emotional and cognitive responses. Although there has been much research conducted on Poe's writing, much of the literature focuses either on biographical summaries or purely on formalistic summaries, and does not relate to the existential themes found in Poe's poems by identifying how the various linguistic structures create the coherent extent of ontological ambiguity that is found in *A Dream Within a Dream*. Thus, this chapter identifies a three-level methodological framework designed to analyze Poe's carefully blended semantic and structural values, and to use a rigorously text-based model to show how each formalistic choice Poe makes also constructs the ontological ambiguity demonstrated in his poem.

1.1 The Research Background

Stylistics is an interdisciplinary subject area within applied linguistics that studies literary works as "texts"; therefore, a literary piece's meanings should be studied in conjunction with how the words of that text are arranged to work together to create a total meaning and the way they create an emotional response when those words are arranged in that method. Within this framework, poetry is at the highest level of study, as poetry contains the most tightly patterned language of any of the literary genres because poets create visual images through their word choices, use the sound of words to play with the meaning of their messages and use line length and stanza design and organization of the text on the page unlike any other poet (Kroll, 2020; Torremocha, 2015). In the current study, stylistics is used as a framework for analyzing Edgar Allan Poe's *A Dream Within a Dream* by showing how the linguistic devices found in it interact with the poem's formal structure to create a very evocative experience of reality, time, and loss. This structure uses language's structure as the primary means of conveying an author's perspective on creation to the reader, giving an existential perspective (Kroll, 2020; Torremocha, 2015).

The primary objective of this analysis is to specify and provide detailed textual evidence of the linguistic features Edgar Allan Poe uses to create both existential uncertainty and affective impact. Through an examination of specific ways in which the poet's diction and imagery choices reflect the poem's thematic concerns, such as transience, illusion, and the dissolution of certainty, this paper has elucidated the manner in which the poet encodes these themes using a lexicon that does not apply a purely bibliographic or biographical frame to the work. This analysis supports an expanded understanding of poetic rhetoric by establishing that sound and meaning influence each other; thus, the prosodic design of language works in tandem with its semantics to create impact. This methodological position stems from the view that rich stylistic description is a necessary part of how tone and mood of a work can be created through linguistics, which positions Poe's lyricism within a rigorous linguistics paradigm and demonstrates how a literary theme alone, without the inclusion of stylistic detail, cannot accurately portray the experience of Poe's writing. (Kroll, 2020; Smith, 1961; Torremocha, 2015).

At various points in Poe's poetry, scholars have approached the analysis of his work from either a biographical or a formalist viewpoint, providing thematic summaries but not employing a consistent approach to language-related devices. While the analysis of Poe's lyric works focuses on consistent themes, including time as a theme, the unreliability of perception as a theme, and the impact of bereavement as a theme, the analysis of these poetic works does not typically apply a linguistic base or methodology to explain how these themes are constructed from a linguistic perspective using the syntactical structure of the poem (Kroll, 2020; Smith, 1961; Torremocha, 2015). Therefore, the examination of *A Dream Within a Dream* contains an advanced stylistic analysis of how Poe's word choice, imagery, rhythm, rhyme, and structural form are coherent and contribute to the experience of ontological anxiety.

Furthermore, this paper analyzes important issues in stylistics concerning how form relates to meaning in poetry. Among those issues is the question of to what extent, if any, one can interpret a poem semantically without also considering its phonological and textual properties. Certain authors interpret Poe's poem with reference to his biography, whereas this paper's stylistic approach examines the poem as a complete, cohesive system in which form and content mutually shape and support one another. Some scholars agree with the author's assertion that Poe's association with a unified poetic theory is warranted, while others question the possibility of decoupling literary form from contextual factors. The current synthesis acknowledges each of these perspectives while contending that adherence to a rigorous stylistic framework provides a more exact analysis of the poem's experiential impact (Kroll, 2020; Torremocha, 2015).

In some empirical evidence, the analysis comprises a three-tiered approach. The first tier of this study includes an inventory of the formal items present in the poem, such as meter, line structure, stanzaic patterns, and syntactic distribution. By mapping the above-noted formal characteristics of the poem, the poem's cadence and perception of time will be illustrated. Secondly, the poem's figurative and tangible elements used to represent these items create an auditory effect that displays Poe's fragmented ontology. The third and final tier of this analysis combines the information provided from the first two tiers into one united explanation as to how Poe appropriates ontological blurriness including the fluidity between what is real versus what is illusory by using an organized method to link semantic significance to a prosodic or formal configuration through every linguistic aspect of the poem (Kroll, 2020; Smith, 1961; Torremocha, 2015).

Overall, this stylistic analysis of *A Dream Within a Dream* serves as a revealing model of how a text-based linguistic approach can clarify aesthetic effects traditionally described as mood or theme. Evidence confirms how objective linguistic data directly link to a coherent and systematic description of literary forms; there are many examples of Edgar Allan Poe's remarkable ability to develop a persistent and vividly meaningful analysis of uncertainty and possibility. Not only does this serve to improve Poe scholarship, through a serious consideration of stylistic methods but also continues to support the position that the study of rhetorical devices in

poetry cannot be separated from an in-depth study of linguistic form, because the success of rhetorical devices in poetry depend upon utilizing phonological, syntactic and semantic materials in a precise manner (Kroll, 2020; Smith, 1961; Torremocha, 2015).

1.2 Gaps

Individual devices have been regularly separated from each other in scholarly analyses of Poe's poetry; however, there has been little attempt to place these analyses together in one coherent, stylistic framework created to distinctly separate meaning, sound, and structure. Analyses of imagery are sometimes not clearly defined regarding which specific senses are revealed by particular imagery, and analyses of sound usually identify only rhyme, without identifying the specific patterns of assonance, consonance, and rhythm that support the meaning. Similarly, references to the poem's formal choices have typically received at best passing reference in the prevailing scholarly discourse and, in general, have not been considered as integral pieces of a common design. Thus, there is a gap in the literature regarding the use of a systematic approach to stylistically read that combines all of these elements into one analytical mechanism and, in turn, makes an explicit connection between each level of language back to the poem's central thematic concept of the impermanent and dream-like nature of being, thus demonstrating how form and content collaboratively create ontological ambiguity. The purpose of this paper is to attempt to fill this gap by situating a unified framework of sense, sound, and structure as mutual resources that drive Poe's investigation of transience and doubt (Aftab et al., 2024; Erjan, 2022; Khan et al., 2023; Stylistic Analysis of Langston Hughes' Selected Poems, 2020).

1.3 Announcement of the Current Research

This research undertakes a full stylistic analysis of *A Dream Within a Dream* by Edgar Allan Poe. After reviewing the author's life and works and considering the relationship between the poem and Poe's biography, the study examines the poem at the levels of sense, sound, and structure. Each level is illustrated with direct textual examples from the poem to demonstrate how Poe's linguistic choices work together to enact, rather than merely state, the poem's anxious meditation on the fleeting and possibly misleading nature of human experience.

2. The Author's Life, Work, and Writing

The biography of Edgar Allan Poe has had an inextricable influence on the development of his literary output, as the early loss of his parents and his adoption into the Allan family created a wild, aspirational, and often difficult course of development that shaped the aesthetic choices and thematic concerns of all of his writing (Napitupulu, 2023; Phillips, 2018; Ulugbekovna, 2020). Poe's family history includes the abandonment by his father and the early death of his mother from tuberculosis, resulting in unsustainable family arrangements culminating in an adolescence with John and Frances Allan, one that was characterized by financial difficulties yet instrumental in his training as a writer and in the development of his first body of literary work (Napitupulu, 2023; Oltean, 2018; Ulugbekovna, 2020). The author's biography aids understanding of his lifelong transition between aspiration and deprivation, as exemplified by his enduring theme of seeking to create innovative forms while exploring ethical ambiguity in his prose and poetry (Bowers, 2010; Tresch, 2016; Ulugbekovna, 2020).

Poe worked as an editor and critic all throughout the northern and mid-Atlantic cities and wrote many of his most famous poems and stories, such as *The Tell-Tale Heart*, *The Raven*, and *Annabel Lee*. His work in these cities reflects a cross-section of American Gothic and dark romanticism, and his influence is seen in the emergence of modern short stories and early detective stories. Additionally, personal loss, economic instability, and competitiveness with his colleagues helped shape the style, rhythm, and psychological depth of Poe's writing.

The underlying principle behind Poe's writings, as most explicitly articulated in his essay "The Philosophy of Composition" (1846), is that the aesthetic experience in poetry is derived from its beauty, while melancholy is an acceptable emotional content for works of poetry. The death of a beautiful woman is considered by Poe to be the ideal subject matter for poetry. The manner in which Poe articulates these aesthetic values can be viewed as a self-reflective construct as well as an explanation of his emphasis on loss and mourning in poetry, connecting his aesthetic choices to an in-depth theory of affect and his artistic structure, which informs the quality of his well-regarded fiction (Meyer, 2013; Pahl, 2009; Pournara, 2022). The connection between Poe's life and emotional state and his critical writings, therefore, helps explain why his poetry and short stories focus on themes of absence, mortality, and the uncanny as structural components of form (Dita, 2019; Pahl, 2009).

As part of their broader scholarly debate on Romanticism, aestheticism, and the evolution of American literary modernism, researchers have incorporated Edgar Allan Poe's life and work into discussions of the latter issues. For example, Poe's biographical/historical pressures and how those pressures were utilized in developing his creative imagination; the cosmopolitan geography of Poe's life; how Poe's experimental/formal developments articulated distinct concepts of doubt and desire; how Poe's writing negotiates boundaries between beauty/horror, self/other, reality/dream (Ke, 2018; Møllegaard, 2016; Pournara, 2022). All these aspects of Poe's

life, work, and writing point toward the importance of a stylistic approach that considers life writing, critical theory, and textual form as three mutually reinforcing elements that shape the way Poe negotiated boundaries between beauty/horror, self/other, and reality/dream (Ivanchenko, 2019; Oltean, 2018; Phillips, 2018). Collectively, the life, career, and writings of Poe reveal a systematic, though contradictory, approach to authorship in which personal experience aligns with cultural audacity to form an enduring legacy of American Gothic modernism (Ivanchenko, 2019; Oltean, 2018; Phillips, 2018).

3. The Poem and the Author

The poem *A Dream Within a Dream* was first published on March 31, 1849, in The Flag of Our Union and is representative of the tumultuous final phase of Edgar Allan Poe's life, which ended with his death in October of that same year. It was written shortly after the passing of Virginia Poe, Poe's beloved wife; at a time when Poe's health and finances were failing; and in a decade characterized by themes of loss and uncertain reality, producing a poem centered on the idea of loss that cannot be undone and questioning whether experience may ultimately be experienced as dreamlike (Kroll, 2020; Smith, 1961; Torremocha, 2015). Many commentators on the poem interpret the "you" — the unnamed addressee of the first stanza — as a representation of a lost love; this is consistent with Edgar Allan Poe's long-held belief that the death of a beautiful woman is the most significant theme for the poet.

At the same time, the poem reflects on bigger ideas about the state we find ourselves in; it reflects on the day and how it is disconnected from the realities of the world we inhabit; and it suggests how the state of being disconnected from the life we have created leads to a disconnection between your identity and the physical manifestations of your true self. By being disconnected from the day or the experiences or the way in which you created your reality, as well as being disconnected from the physical manifestations of your existence, you lose all sense of reality and become subject to your own thoughts of what could be if only you could achieve your dreams. By including this distinction in the author's relationship with the poem, the author creates an opportunity to shift their grief from an individual experience to a philosophical consideration by examining death in relation to the nature of existence. By doing this, the writer also aligns with Edgar Allen Poe's statement that poetry must seek beauty and heightened emotion rather than biographical insight (Kroll, 2020; Smith, 1961; Torremocha, 2015).

The circumstances surrounding the poem's author will not provide an interpretation of the poem. As a result, by refraining from using specific names, places, and events, Poe encourages readers' access to implicit connections between affective experience and existential doubt. The tension between personal intention and universal commitment is a hallmark of Poe's stylistic method, thereby strengthening the relationship between form and feeling and creating a powerful dramatic effect. Critics who interpret the poem about Virginia's death informally connect how one experiences personal loss through public expression, yet the poem's stylistic interpretation emphasizes how its use of sound, word choice, and grammatical structure creates an interactive expression of uncertainty as well as a dynamic experience of existence. In addition to this outcome of double meaning, it is possible to achieve these forms of meaning within the formal structure of the poem (Aftab et al., 2024; Erjan, 2022; Khan et al., 2023).

As part of the general field of Poe studies, the poem *A Dream Within a Dream* represents a confluence between textual form, life-writing, and critical theorizing. As an instance from Poe's adult life, the way he composed this poem later in his life serves as a case study in how Poe's ongoing engagement with loss, memory, and the uncanny manifests in formal devices such as diction, imagery, and prosody. As such, the wider academic discussion about this poem has benefited from a stylistic perspective that treats life-history not just as a biographical note but as part of a broader, disciplined reading of how poetic formal devices articulate experiences of absence and the irreversibility of time. This analysis places a strong emphasis on the manner in which the poet has converted privately experienced grief into a universal meditation on the nature of perception and temporality; however, at the same time it acknowledges that the life of the poet is only one of many important factors in this work's aesthetic autonomy and its ability to depict the outside world (Napitupulu, 2023; Phillips, 2018; Stylistic Analysis of Selected Poems by Langston Hughes, 2020).

4. Research Methodology

A qualitative text image study that uses grammatical structure, sound patterns, and semantic choices in combination to create meanings for literature. Using causal history and isolated form critique, this methodological approach applies the text of Poe's *A Dream Within a Dream* as a complete system of interdependent constituents. The development of stylistics as an interdisciplinary point of reference for this methodology, and its relationship to poetic form and artistic effect, lends credence to the empirical design of this research. The analysis of textual data will enable mapping production to the poet's aesthetic choices based on shared qualitative descriptions.

A formal structural inventory of the text facilitates a full description of the poem using three distinguishing tiers. The first level is an inventory of the poem's structural and quantitative characteristics, including lineation (placement), stanzaic patterning, and the synchronous presentation of syntax and meter. This descriptive level isolates the poem's mechanical architecture: how structural cadences govern the reader's perception of time and pace through the placement of lines within stanzas.

Tier two of the framework shifts the analytical focus to the overlapping levels of sense and sound in the poem. On the sense level, this tier analyzes the use of specific figures of speech (e.g., metaphor, personification), semantic weight, and sensory imagery (i.e., tactile, visual, auditory, kinesthetic). At the same time, this tier analyzes the poem at the phonological level to identify and categorize sound imagery devices, including end rhymes, internal rhymes, assonance, consonance, and alliteration, and to examine how these sound devices create or reinforce shifts in the poem's mood.

Tier four synthesizes the data through the preceding three tiers to create a consistent linguistic rationale. Specifically, by tracing how the semantic significance of an experience intersects with its prosodic and formal structure, tier four describes how Poe's poetry takes a personal experience of grief and transforms it into a universal/philosophical meditation. This example of Poe's grief demonstrates how the "fluid, unstable boundary" between reality and illusion (the poem's central ontological ambiguity) is "actively created" through specific phonological, syntactic, and semantic choices rather than merely "stated."

5. Poetry Analysis

In this chapter, the author conducts a thorough, three-part analysis of Edgar Allan Poe's *A Dream Within a Dream*, divided into the three categories of sense, sound, and structure. The sense analysis illustrates how images of touch, sight, and sound provide a grounding for emotions related to existential dread/uncertainty through metaphors, personification, and rhetorical framing. The sound analysis shows how Poe's use of rhyming couplets, iambic meter, and assonance, consonance, and alliteration replicates the breakdown of the "real" by using musicality and rhythm to portray emotions. The structure of *A Dream Within a Dream* is evident in irregular line lengths, asymmetrical stanzas, and enjambment, which mirror the instability of time and the human experience.

5.1 Sense

5.1.1 Figures of Speech

Poe develops the poem with a limited number of similar terms or phrases. There is an extended metaphor, maybe 6 or 7 examples of personification, and each of the three rhetorical devices used at the close can also be viewed as a type of refrain. The controlling metaphor and metonymy express the theme of loss in the original text's metaphor for loss of life or love, as expressed in lines 14-16: "Grains of the golden sand— / How few, but how they creep / Through my fingers to the deep." In these lines, the metaphor offers a concrete way to think about lost time. It creates a link to the concept of "lost time" as described through the metaphor expressed in lines 10-11 — repeated at the poem's close — and concludes with the statement, "all we see or seem / Is but a dream within a dream." Thus, the speaker connects all existential experiences as if they are nested dreams, creating an additional layer of unreality between the speaker and any stable truths.

To heighten the emotional significance of this unreality, the writer employs personification to make "hope" (the abstract thing) seem as though it can fly away from the speaker (line six). When hope is described as having wings and being alive, it heightens the feeling of loss by shaping the reader's perception of how something that is merely an abstract idea (hope) can actually leave the speaker. The way hope leaves the speaker corresponds with the use of symbolism in the environment by the "surf-tortured shore" and the "pitiless wave", where the surf-tortured shore and pitiless waves represent the continuous, unfeeling passage of time and death. The shore is, in essence, the intangible boundary of human life, constantly eroded by forces beyond one's control. While the poem does not offer definitive answers to these themes, it poses rhetorical questions such as "Is it therefore less gone?" and "Is all that we see or seem / Just a dream within a dream?" The use of these questions illustrates the speaker's own uncertainty and, in turn, invites the reader to experience the same epistemological uncertainty as the narrator.

5.1.2 Imagery

Sound is not widely present in the poem, with only one strong occurrence, while the imagery is largely limited to touch and sight. The tactile imagery's hold on the reader creates physical disappointment through the line "I have in my hand / Grains of the golden sand— / How many! Yet how do the grains / Slip between my fingers." Hence, the tactile disappointment parallels the narrator's inability to hold on to time or their love, as the sand is lost. The tactile frustration is visually identified through two contrasting images: the author begins with a close-up image of departure, "Take this kiss upon the brow," and moves toward the opposite end of the spectrum, the expansive shoreline scene with golden sand.

The auditory imagery of “the roar / Of a surf-tormented shore” provides an enhanced setting for the statement, as it not only heightens the sound of the sea crashing onto the beach but also reinforces both the violence and inevitability of the natural forces opposing the speaker. Finally, kinesthetic imagery ties the two together with “while they creep / Through my fingers to the deep.” By conveying a continuous sense of moving forward at an extremely slow, irreversible pace, the narrative’s time frame has successfully portrayed time as something to be physically experienced rather than merely thought of in the abstract.

5.1.3 Denotative and Connotative Meanings (Semantic Level)

In terms of semantics, the reader finds that many words in this poem have strong connotative indicators that go far beyond merely defining them literally or in a dictionary. For example, the word “dream,” when viewed denotatively, refers to thoughts/experiences you have while sleeping. However, when thinking of the word “dream” in that context and then applying it to all of what is not real, or has no substantiating evidence of truth (to life upon waking, etc.), its connotation expands to encompass several other ideas. Likewise, the phrase “golden sand” is also an example of extending the denotative definition of a word into another category. When looking at the denotative definition of “golden sand,” sand particles found along a beach are the first thing that comes to mind. When the adjective “golden” describes the noun “sand,” the adjective causes you to think of something that is probably of a precious value, and therefore, the sand becomes symbolic of precious and finite memories in a person’s lifetime.

The elements, as they appear in nature, have also been affected semantically. For example, looking at how the term “pitiless wave” is a denotation of something moving in a normal, everyday manner (i.e., ocean waters), but the application of the word “pitiless” creates a characterization of the sea as being heartless and cruel; therefore, the sea has now become an example of a non-characteristic, yet powerful agent that acts like time or death. In conjunction with that example, the phrase “surf-tormented shore” has been impacted in the same manner. A purely denotative definition of the surf-tormented shore describes the shoreline as an area of the beach that is consistently washed by ocean waves. However, by using the word “tormented,” the intense suffering that is typically felt or experienced by humans can now be similarly identified with the physical characteristics of the shoreline.

5.2 Sounds

Poe uses sound devices such as end rhyme, internal rhyme, rhythm, assonance, consonance, and alliteration to create the poem’s emotional and philosophical meaning. The poem’s rhyming pattern is almost exclusively based on couplets and AABB-type rhymes (e.g., “brow” / “now”, “deem” / “dream”), resulting in a sound pairing that gives the poem an incantation-like quality that suits a person caught in obsessive, repetitive thoughts. The poem’s structural use of a repeated line (the speaker’s last line of the poem is the same as the first two lines of the second stanza) repeats the same cycle as the repeated line occurs at the end of the first stanza. The circularity of the text establishes a cycle of haunting/unresolved themes and a cycle to the end of the work.

The rhythm of the poem controls its movement. The majority of the poem’s lines have an iambic meter and vary in syllable count, producing a rocking, wave-like rhythm that simulates how the surf moves as described in the text. In this case, the long “o” and “e” sounds created by repeated assonance in “How few! yet how they creep” slow the line down, mimicking the gradual retreat of sand left behind by the waves. In addition, the same “s” sound in the phrase “surf-tormented shore” illustrates consonance and creates a hissing effect that mimics waves hitting land.

The sibilant quality of this sound carries into the poem through alliteration, particularly with the word “see” in the phrase “see or seem.” In this case, the repetition of the same initial letter (“s”) connects both verbs, linking them phonetically as well as conceptually, creating an ambiguity between reality and appearance for the reader through word choice. Overall, the sonic landscape created by these literary devices culminates in many instances of repetition and anaphora-like repetitions, including but not limited to “While I weep — while I weep!” Each occurrence has a heightened emotional effect, illustrating the speaker’s sorrow with sound rather than simply by writing about it.

5.3 Structure

This poem has a total of twenty-four lines, divided into two stanzas, each varying in length from eleven to thirteen lines. The line lengths themselves are irregular; they range from approximately four to eight syllables, resulting in a halting, panting rhythm rather than a steady one. Most of the lines in the poem are grouped in pairs that rhyme (couplets), with just a few exceptions where three lines rhyme (triplets), such as in lines 18 and 19, as well as their surrounding rhymes on “grasp” / “clasp” and “save” / “wave.”

The grammatical form of these lines is not always whole, complete sentences; some of them run on to the next line (enjambment). For instance, the meaning of “Yet if hope hath flown away / In a night, or in a day” will not be resolved until you finish reading line seven. The continuing overflow of meaning at line breaks parallels the

central image of all things within the poem — hope, sand, time — disappearing consecutively and in greater amounts without ever being contained within a defined boundary. The division into two stanzas also represents a transition in both place and address — the first stanza addresses an unnamed "you," connected with a moment of separation, whereas the second stanza addresses God instead of person, and includes a significant transition from a moment of intimacy associated with the person being left to being at sea on the shore. The break in structure at the level of form is the same as the breaking of the poem's theme; from a specific loss of human contact to an abstract or philosophical discussion about the uncertainty of the external world.

6. Conclusion

The linguistic choices that Edgar Allan Poe makes at every level — sense, sound, and linear structure — within *A Dream Within a Dream* are deliberately designed to enact his central concern within the poem: that love, hope, time, and life itself may essentially be no more real than a dream. In terms of sense, Poe uses the extended metaphor of sand slipping through one's fingers to give the abstract philosophical anxiety expressed in the poem a concrete, physical manifestation. This is accomplished through tactile and visual imagery and the use of words having other than or additional meanings. In terms of sound, the use of couplet rhyme, iambic meter with a wave-like rhythm, and the use of assonance, consonance, and alliteration create, almost physically, the ceaseless movement of the ocean and the ceaselessly repeated, obsessive nature of grief.

The poem's two unequal stanzas, varied line lengths, and frequent use of enjambment at the structural level create a form of instability and ungraspability that is consistent with the poem. In the context of Poe's numerous biographical losses, along with his professed beliefs that melancholy and the death of a lost female are the most appropriate themes for poetry, the theme of personal sorrow has been artistically crafted through style to become a universal expression of the transitory nature of human perception and the inevitability of temporal loss.

References

Aftab, M., Shabir, M., Bushra, B., Khan, D., & Khan, H. (2024). A Guide to Stylistic Analysis: A Case Study of Daud Kamal's "An Ode to Death." *Qlantic Journal of Social Sciences and Humanities, 5*(1), 248–260. https://doi.org/10.55737/qjssh.194649332

Bowers, J. S. (2010). Chasing Edgar: The Tourist Rhetoric of the Poe Bicentennial. *Poe Studies, 43*(1), 59–82. https://doi.org/10.1111/j.1754-6095.2010.00026.x

Dita, P. D. (2019). A Late Victorian Interference of Genres: Aestheticism Reshaping the Faust Myth in the Picture of Dorian Gray. *Humanitas - Uluslararası Sosyal Bilimler Dergisi, 7*(14), 337–357. https://doi.org/10.20304/humanitas.563477

Erjan, R. Sh. A. (2022). A Stylistic Analysis of John Keats's Poem "Ode to Psyche." *Theory and Practice in Language Studies, 12*(7), 1378–1385. https://doi.org/10.17507/tpls.1207.18

Ivanchenko, M. Yu. (2019). *REPEATITION AS A STYLISTIC DEVICE OF EXPRESSION IN E. PO POETIC SPEECH*, 109–124. https://doi.org/10.36059/978-966-397-131-5/109-124

Ke, X. (2018). The Influence of Edgar Allan Poe's Life Experience on His Writings from the Psychoanalytic Perspective. https://doi.org/10.2991/icassee-18.2018.18

Khan, A. A., Khan, I. A., & Gul, N. (2023). Exploring Stylistic Elements in Daud Kamal's Poem Plough & Oxen. *Journal of Policy Research, 9*(4), 318–322. https://doi.org/10.61506/02.00154

Kroll, S. (2020). La sonoridad de la culpa, malos presagios y escenas de horror: a propósito de la asonancia calderoniana en u-a. *Revista De Literatura, 82*(164), 491–514. https://doi.org/10.3989/revliteratura.2020.02.019

Mazurkiewicz, M. (2018). Edgar Allan Poe: A question of sport. *Literatura I Kultura Popularna, 23*, 199–206. https://doi.org/10.19195/0867-7441.23.9.14

Meyer, J. (2013). The Marian Aesthetics of Edgar Allan Poe. *The Edgar Allan Poe Review, 14*(1), 1–15. https://doi.org/10.5325/edgallpoerev.14.1.0001

Møllegaard, K. (2016). Fame after Life: The Mystery of Edgar Allan Poe's Death. *Folklore Electronic Journal of Folklore, 65*, 129–150. https://doi.org/10.7592/fejf2016.65.mollegaard

Napitupulu, R. (2023). THEMES IN SOME POEMS BY EDGAR ALLAN POE. *Linguistik Jurnal Bahasa Dan Sastra, 8*(4), 628. https://doi.org/10.31604/linguistik.v8i4.628-639

Oltean, R. (2018). Philip Edward Phillips (ed.), Poe and Place. *Transatlantica*, (2). https://doi.org/10.4000/transatlantica.13295

Pahl, D. (2009). Sounding the Sublime: Poe, Burke, and the (Non)Sense of Language. *Poe Studies, 42*(1), 41–60. https://doi.org/10.1111/j.1947-4697.2009.00012.x

Phillips, P. E. (2018). Poe and Place. https://doi.org/10.1007/978-3-319-96788-2

Pournara, L. (2022). John Michael, Secular Lyric: The Modernization of the Poem in Poe, Whitman, and Dickinson. *European Journal of American Studies*. https://doi.org/10.4000/ejas.15666

SMITH, C. C. (1961). La musicalidad del "Polifemo." *Revista De Filología Española, 44*(1/2), 139–166. https://doi.org/10.3989/rfe.1961.v44.i1/2.992

Stylistic Analysis of Langston Hughes' Selected Poems. (2020). *Zanco Journal of Humanity Sciences, 24*(1). https://doi.org/10.21271/zjhs.24.1.16

Swarnakar, S. (2017). REPRESENTATION OF DEATH IN EDGAR ALLAN POE AND EMILY DICKINSON. *A Cor Das Letras, 8*(1), 29. https://doi.org/10.13102/cl.v8i1.1563

Teitelroit, R. C. (2020). Poe's Raven says more than "nevermore." *Aquila*, (23), 201–210. https://doi.org/10.17648/revista-aquila.v1i23.124

Torremocha, M. V. U. (2015). RESEÑA de: Carandell, Zoraida (ed.): Traduire pour l'oreille. Versions espagnoles de la prose et du théâtre poétiques français (1890-1930). París: Presses Sorbonne Nouvelle, 2014. Rhythmica Revista Española De Métrica Comparada, (13), 235. https://doi.org/10.5944/rhythmica.16160

Tresch, J. (2016). "Matter No More": Edgar Allan Poe and the Paradoxes of Materialism. *Critical Inquiry, 42*(4), 865–898. https://doi.org/10.1086/687203

Ulugbekovna, R. S. (2020). Historical Depiction of Great Person "Tamerlane" In the Poem of Edgar Allan Poe. *The American Journal of Social Science and Education Innovations, 02*(07), 123–128. https://doi.org/10.37547/tajssei/volume02issue07-14

Paradigm Academic Press
Art and Society
ISSN 2709-9830
JUN. 2026 VOL.5, NO.2

Research on the Unique Value of Wall Painting Art as a Non-Literal Carrier

Guohong Ran[1]

[1] Lingnan Normal University, Guangdong, China

Correspondence: Guohong Ran, Lingnan Normal University, Guangdong, China.

doi:10.63593/AS.2709-9830.2026.06.002

Abstract

As an immersive visual art deeply embedded in urban and rural public spaces, wall painting art breaks through the limitations of text carriers of expression and takes intuitive visual symbols, scene immersive experience, and public interactive attributes as its core features, becoming a very dynamic contemporary non-textual culture and emotional carrier. This article takes wall painting art as the core research object, relying on cultural memory theory, non-verbal communication theory and spatial narrative theory to analyze its differentiated advantages compared with traditional text carriers, and deeply explore its role in individual emotional comfort, collective memory cohesion, regional context inheritance, and urban and rural space. Activate the unique value of the four dimensions, sort out the practical difficulties existing in the current wall painting creation and application, such as homogeneity, superficiality, lack of long-term operation and maintenance, and isolation from public needs, and propose optimized paths for localized creation, connotative expression, normalized operation and maintenance, and universal participation. The research aims to improve the theoretical system of non-literal cultural carriers and provide theoretical support and practical reference for the construction of urban and rural public culture, the living inheritance of nostalgic culture, and the innovative application of public art.

Keywords: wall painting art, non-textual carrier, cultural memory, collective emotion, cultural inheritance, space activation

1. Introduction

The inheritance of cultural memory relies on specific carriers. The form, attributes and communication capabilities of the carrier directly determine the vitality and breadth of regional cultural continuation. For a long time, written carriers have been the core medium for recording regional culture, preserving collective memories, and inheriting nostalgic emotions. Textual materials such as local chronicles, literary and historical classics, folk documents, and handwritten archives have systematically preserved the evolution of regional history, folk customs, and social changes, and are an important basis for the continuation of humanistic context. However, in the contemporary communication context, the limitations of traditional text carriers have become increasingly prominent. It relies on rational reading and logical interpretation of communication methods, which places high demands on the audience's cultural literacy and cognitive level, and is difficult to adapt to the public cultural communication needs of the entire population. In the new media era, the public is more adapted to the fragmented, visual, and scene-based information reception mode. Abstract and solid text communication is gradually out of touch with public life, which in turn leads to practical problems such as fragmentation of collective memory, faulty folklore inheritance, and the dissolution of urban and rural cultural characteristics. As a typical visual art rooted in urban and rural public spaces, wall painting uses building walls, street facades, and idle enclosures as carriers to transform abstract cultural connotations such as regional historical stories, folk customs, and local life images into perceptible and visible immersive visual scenes. It is a typical space-embedded non-text cultural carrier. (Chen Li & Liu Yang, 2024) Compared with art forms such as

painting, photography, and sculpture, wall painting has the advantages of publicity, openness, locality, and low cost. It can be deeply integrated into people's daily life and realize the normalized spread of cultural memory and local emotions.

2. Literature Review and Research Status

Since the 17th century, academic circles have begun to pay attention to the relationship between nostalgia and memory carriers. Early studies defined nostalgia as pathological homesickness. Subsequently, with the cross-development of sociology, cultural studies, and psychology, nostalgia was redefined as a positive collective cultural psychology and identity construction method (Johnson S., 2018). At the level of carrier research, research has long focused on material carriers such as architectural heritage, cultural relics, and traditional handicrafts, as well as non-material carriers such as folk rituals and behavioral customs. It is believed that non-textual carriers take sensory experience and emotional resonance as the core communication logic and can break through the threshold of written cognition and realize the popular communication of cultural memory (Brown D. & Davis J., 2019). In recent years, the research perspective has gradually extended to modern visual media. Many scholars have explored the memory-carrying function of film, television, photography, and public installation art. However, systematic research on the special carrier of wall painting is still relatively lacking. The study of wall painting art originated from the study of street graffiti culture. In the early days, it focused on the social expression and personality release of street art, and then gradually turned to the study of public space reconstruction and community cultural construction. (Miller T., 2020) Research has confirmed that community wall paintings can record the process of spatial change, build emotional attachment between residents and the region, and strengthen the collective memory of the community. (Smith L., 2017) Taking community wall paintings in American cities as a research sample, it is proposed that wall paintings have a public narrative function, which can realize the organic integration of individual emotional memory and community collective memory, and promote the living inheritance of regional culture. Overall, the research has fully verified the cultural and emotional value of wall paintings, but the research scenarios are mostly limited to urban communities, and targeted research on rural areas, local nostalgia, and folk custom inheritance is relatively weak.

Comprehensive review of existing research shows that there are still three core research gaps in the current academic circle: First, the carrier research perspective is solidified, focusing on literal carriers and traditional material carriers for a long time. There is insufficient theoretical research on modern non-literal public visual carriers such as wall paintings, and no corresponding carrier analysis system has been constructed; second, the research on the value of wall paintings has a single dimension. Most of them focus on surface practical values such as landscape beautification and space activation, and ignore their deep core values of emotional carrying, memory reconstruction, and cultural context continuation; thirdly, there is a lack of comparative research, and the essential differences and competitive advantages of wall paintings and traditional text carriers and other visual carriers are not systematically clarified, and the irreplaceability of wall painting carriers is insufficiently demonstrated.

3. Core Concepts and Differentiated Characteristics of Carriers

The non-textual carrier of wall painting art defined in this article is a public art medium that relies on urban and rural building walls, street facades, public enclosures and other physical spaces, uses hand-painted visual images as the core expression form, focuses on regional historical changes, folk customs, local life and collective emotions, and can independently complete cultural narrative, memory retention and emotional transmission without the aid of text (Wu Jie, 2023). Different from the static text attributes of traditional text carriers, wall paintings are a typical space-embedded living cultural carrier. They can break through the time and space constraints and cognitive limitations of text communication, transform abstract regional culture and invisible nostalgia into visual scenes solidified in public space, and promote the deep integration of cultural memory, collective emotions and urban and rural physical spaces. They are a new non-text carrier for contemporary public cultural communication.

In order to systematically clarify the unique competitive advantages of wall paintings as non-literal carriers, this article differentiates them from traditional literal cultural carriers from the five core dimensions of expression form, cognitive threshold, communication scene, emotional expression and iterative update. The specific characteristics are as follows: In terms of expression form, traditional text carriers rely on abstract text symbols and logical narratives to complete cultural records, express indirect generalizations, and require the audience to imagine and restore the emotions of the scene. Wall paintings, on the other hand, use concrete images and scene reproductions as their core expressions to intuitively present regional features and avoid the cumbersome interpretation of text. For example, the farming-themed wall paintings created in many rural areas in China directly present local scenes such as rice transplanting, autumn harvest, and ancient sericulture through hand-painted pictures (Figure 1). Without the need for elaboration, the audience can intuitively understand the

connotation of local farming culture, and the information transmission is efficient and intuitive (Guo Xin, 2021).

Figure 1. Harvest Abundance

In terms of cognitive threshold, text carriers have high requirements on the cultural level and logical thinking of the audience, and there are obvious barriers to communication circles. Wall painting is suitable for people of all ages and levels, and has the advantage of zero-threshold communication. Neighborhood cultural wall paintings in old urban communities and rural folk custom wall paintings allow children, elderly groups, and foreign tourists to perceive cultural content through visual images, effectively breaking the cognitive limitations of textual communication and realizing the universal dissemination of regional cultural memory. In terms of communication scenarios, text carriers are mostly sealed in libraries and archives, separated from daily life, and communication is passively solidified. Wall paintings are deeply integrated into public spaces to achieve normalized immersive communication. Various theme wall paintings on the walls of rural streets, urban community fences, and street leisure areas are completely embedded in people's daily travel, leisure, and social life, so that cultural communication is no longer limited to book reading, and cultural infiltration and memory transfer can be completed anytime and anywhere (Huang Bo, 2022). In terms of emotional expression, text carriers focus on rational recording of historical facts, and emotional expression is introverted and weak, making it difficult to form deep resonance. Wall paintings rely on artistic creation to impart humanistic warmth to the space. For example, wall paintings with nostalgia themes in many places accurately capture the public's nostalgic emotions by replicating life-like scenes such as fireworks in old alleys, neighborhood chats, and childhood play, awakening the growth memories and local feelings of local residents, and achieving immersive emotional experience and spiritual resonance (Figure 2). In terms of iterative updates, the content of traditional text carriers is solidified and difficult to modify, and it is easy for cultural inheritance gaps to occur. Wall painting has the advantage of flexible and dynamic iteration, and the creative content can be updated with the times. On the basis of retaining the core of traditional farming and folk culture, many villages have renovated their walls and added modern theme wall paintings such as rural revitalization, ecological livability, and new farmer life to achieve the integration of traditional local culture and new era development concepts, allowing regional culture to innovate in inheritance and adapt to the development needs of the times (Chen Li & Liu Yang, 2024).

Figure 2. Childhood

4. The Multidimensional and Unique Value of Mural Art as a Non-Verbal Medium

4.1 Individual Dimension: Emotional Redemption and Spiritual Solace

In contemporary society, characterized by rapid urbanization and large-scale cross-regional population movements, mass migration has disrupted traditional rural lifestyles. A vast number of migrant workers and young students have been separated from their hometowns for extended periods, resulting in a prolonged disconnect from their native rural environments, folk customs, and neighborhood relationships, which has given rise to widespread spiritual nostalgia and a sense of spatial alienation (Yang Lin, 2020). At the same time, modern urban spaces suffer from severe homogenization; standardized architectural layouts, fast-paced lifestyles, and an atmosphere of interpersonal indifference have plunged contemporary people into spiritual emptiness and emotional anxiety, leaving them lacking a stable sense of belonging and a source of solace. Although traditional textual media can document rural history and life memories, they only preserve memories at the level of rational cognition; they cannot alleviate concrete emotional anxiety and struggle to meet the public's need for spiritual solace. As a non-textual visual medium, mural art possesses unique value in emotional redemption and spiritual solace, serving as an important medium for alleviating modern nostalgia and easing spiritual anxiety. Mural art focuses on the public's most profound, everyday memories. Through hand-painted imagery, it recreates concrete scenes such as rural farming labor, traditional folk festivals, the atmosphere of old street markets, casual interactions among neighbors, and childhood scenes, transforming the public's vague, fragmented, and abstract feelings of nostalgia into clear, visible, and immersive visual scenes (Li Yue & Wang Qiang, 2022). For those living far from home, these rural murals in the streets and alleys construct a tangible "visual homeland" that can quickly awaken deep-seated memories of the countryside and fill the emotional void caused by physical separation; for local residents, these familiar scenes evoke memories of growing up and alleviate the sense of alienation and loss brought about by the spatial changes of modernization. Compared to cold, objective textual records, mural art possesses a rich human warmth. Mural creation is not a mechanical replication of scenes, but rather an artistic reinterpretation that integrates the artist's understanding of the local landscape, the collective memory of the public, and the cultural heritage of the region, with every detail resonating with the public's subjective emotional experiences (Zhao Lei, 2023). This immersive, warm, and accessible method of emotional communication is a core advantage that textual media cannot replace. It effectively alleviates the spiritual anxiety of contemporary people, provides individuals with a stable emotional anchor and a sense of spiritual belonging, and achieves emotional redemption and psychological healing for the public.

4.2 Collective Dimension: The Value of Collective Memory and Community Identity

The cohesion and sense of belonging within a local community stem from a shared collective memory and cultural identity; the preservation and transmission of collective memory are the core foundation for maintaining

emotional bonds within the community and building a sense of community identity (Han Min, 2021). Individual memories are scattered and fragmented, making it difficult to form a collective consensus. Meanwhile, traditional textual media often record local history in official, professional, or academic formats; while their content is rigorous, it lacks public interactivity. Consequently, they struggle to evoke widespread emotional resonance and fail to effectively integrate the public's scattered individual memories, leading to the gradual fragmentation and dilution of collective local memory.

As a public, non-textual medium rooted in community public spaces and open to all community members, murals possess tremendous value in consolidating collective memory and shaping community identity. The creative materials for murals all stem from the shared life experiences, local folk customs, and collective growth memories of the local community. By circumventing the technical and exclusive nature of textual narratives, murals can accurately capture the shared nostalgic emotions and cultural perceptions of the local community (Wang Shuo, 2023). Unified regional visual symbols, standardized local scenes, and shared folkloric imagery can integrate, synthesize, and unify the fragmented and diverse individual memories of community residents, thereby constructing a complete, systematic, and shared system of regional collective memory. This approach resolves the issue of fragmented memory and strengthens the community's emotional consensus.

4.3 Cultural Dimension: The Value of Preserving Regional Cultural Context and Revitalizing Culture

Regional cultural context is the spiritual core of urban and rural development. Composed of local folk customs, historical narratives, traditional crafts, rural landscapes, and humanistic spirit, it serves as the defining characteristic that distinguishes the cultural traits of different regions (Zhang Yuan, 2022). Amid the rapid transformation of urban and rural spaces, a vast number of traditional folk customs, rural scenes, and cultural practices are gradually disappearing. Regional culture is confronted with the challenges of a severed cultural continuity, erosion of unique characteristics, and a disruption in its transmission. Although traditional textual media have comprehensively documented the core content of regional cultural heritage, they are constrained by high barriers to access, limited reach, and rigid formats. As a result, they struggle to facilitate the widespread dissemination of local culture to the general public or ensure its transmission to younger generations. This has led to many regional cultures being preserved solely in historical texts and classical works, unable to integrate into everyday life or regain vitality in the modern era. As a visual and dynamic non-textual cultural medium, mural art effectively addresses the practical challenges of preserving regional cultural heritage, facilitating the visual dissemination, popular transmission, and innovative development of traditional culture. On the one hand, murals transform abstract textual contexts, obscure folklore theories, and intangible humanistic spirits into intuitive and vivid visual imagery. This allows the public to subtly perceive local culture, understand regional history, and become familiar with folk customs during their daily walks and leisure activities, significantly lowering the barriers to perceiving and learning about regional cultural heritage (Liu Hao, 2024). Compared to the passive, reference-based transmission of written texts, murals have made cultural transmission a regular, proactive, and everyday part of life, bringing the dormant culture of classical texts out of the pages and into the public sphere. On the other hand, murals possess a strong capacity for contemporary innovation, enabling the creative transformation and innovative development of traditional cultural heritage. While upholding the core of local culture, creators combine contemporary aesthetic trends, elements of modern life, and progressive development concepts to offer artistic and innovative interpretations of traditional folk customs and rural scenes. This breaks down outdated stereotypes of traditional culture and aligns with the aesthetic preferences of younger generations. Rural mural art in Guangdong, China, is particularly representative. These works not only fully preserve the core essence of traditional agricultural culture and local folklore—recreating the authentic rural landscape—but also incorporate contemporary elements such as rural revitalization, modern farmers, and eco-friendly villages. This achieves an organic fusion of traditional cultural heritage and modern development, allowing regional cultural traditions to maintain their vitality and be reborn in the new era. At the same time, the open nature of public spaces allows the regional culture embodied in these murals to be widely disseminated to tourists and visitors from outside the area, effectively expanding the influence and reach of local culture and sustaining the humanistic heritage of both urban and rural areas (Figure 3).

Figure 3. Friends

4.4 Spatial Dimension: The Value of Revitalizing Gray Spaces and Humanistic Empowerment

Traditional textual cultural media serve only to record culture and transmit emotions; they cannot participate in the transformation of urban and rural spaces and lack the capacity to empower them. In contrast, as a non-textual medium embedded within space, murals are deeply integrated with the physical environments of both urban and rural areas. They can effectively revitalize idle "gray spaces" in these settings, infusing cold urban architecture and monotonous rural walls with humanistic significance, thereby achieving the dual value of beautifying the spatial landscape and fostering cultural empowerment (Jiang Yu, 2024). In the process of urban and rural development, a vast number of marginal spaces—such as idle rural walls, urban community perimeter walls, construction site hoardings, and the walls of old alleys—have long remained idle, derelict, and monotonous. Lacking cultural vibrancy and aesthetic appeal, these areas have become weak links in urban and rural public cultural development. Wall mural art precisely addresses the transformation needs of these peripheral "gray spaces," achieving artistic and cultural upgrades through its advantages of low cost, high adaptability, and strong creativity. Through hand-painted creations, these once-dull, cold, and worthless blank walls are transformed into cultural showcases that embody local memories, regional culture, and the spirit of the times. This not only enhances the landscape and character of urban and rural spaces and addresses visual shortcomings in public spaces but also endows these idle spaces with profound humanistic significance and cultural functions (Zhou Yang, 2023). This spatial revitalization is not merely landscape decoration but a deep integration of culture and space, transforming urban and rural public spaces from mere physical venues for activities into humanistic spaces that embody aesthetic, cultural, and memorial value. At the same time, the spatial integration of murals allows nostalgic culture and regional cultural context to break free from the closed transmission model of traditional textual media, fully integrating them into the public's daily living spaces and achieving the spatial embodiment and sustained transmission of cultural memory (Figure 4). Through spatial empowerment, murals establish a virtuous cycle of interaction among "space, culture, and people," enabling the public to perceive culture, awaken memories, and foster emotional connections through spatial experiences, thereby continuously revitalizing the cultural vitality and humanistic charm of urban and rural public spaces.

Figure 4. The Passage of Time

5. Current Challenges in the Application of Mural Art as a Non-Textual Medium

5.1 The Proliferation of Template-Based Creations and the Loss of Local Cultural

Characteristics as mural art rapidly gains popularity in urban and rural public projects, issues of homogenization and template-based creation have become increasingly prominent, severely undermining the value of mural art as a non-textual medium for preserving regional memory (Xu Peng, 2023). Currently, most urban and rural mural projects prioritize rapid implementation and cost reduction as their core objectives. Creative teams often mass-produce works using generic online templates, blindly applying popular elements such as pastoral landscapes, ancient-style alleys, and generic folk customs, while failing to conduct in-depth research and exploration of local regional culture, unique collective memories, and distinctive folk traditions. Mural content across different regions, towns, and communities is highly similar, resulting in a widespread phenomenon where "every wall looks the same" and "every village has the same scenery." The core value of murals lies in their locality and uniqueness—that is, their ability to convey a region's distinctive collective emotions and cultural context by drawing on its own historical memories, folk culture, and scenes of daily life. Template-based creation completely strips wall murals of their local cultural essence, reducing them from non-verbal cultural carriers of regional memory to superficial landscape decorations devoid of substance, distinctiveness, or memory. Homogenized imagery fails to evoke local residents' unique nostalgic memories or elicit deep emotional resonance, thereby completely undermining the core value of wall murals as non-verbal cultural carriers and severely hindering the differentiated transmission of distinctive regional cultures.

5.2 Superficial Creation Is Widespread, and Cultural and Emotional Depth Is Shallow

Currently, a large number of mural creations suffer from the problem of prioritizing form over substance. Creators excessively pursue visual aesthetic effects while neglecting the exploration of cultural significance and the shaping of emotional value, resulting in a severe lack of cultural depth and emotional resonance in murals as a medium (Fan Ze, 2024). Most creative teams focus solely on the visual aesthetics of color, composition, and form, simply piling on folk symbols, rural elements, and retro scenes to achieve only a superficial visual recreation, without delving into the historical origins, cultural symbolism, folk connotations, or emotional depth behind the images. This superficial creative approach results in a severe disconnect between mural imagery and

local cultural memory as well as the public's genuine emotions. Although the images appear to align with the theme of rural nostalgia, they lack cultural depth and emotional integration. Consequently, they fail to convey the core spirit of the regional cultural context and struggle to evoke the public's deep-seated memories of growing up and feelings of nostalgia. Ultimately, such murals offer only a fleeting visual experience; they fail to fulfill their core functions of cultural transmission, emotional solace, and the consolidation of collective memory. They are reduced to mere formalistic artistic decoration, unable to fully realize the cultural and emotional value inherent in non-textual media.

5.3 Lack of Maintenance Mechanisms Makes the Preservation of Mural Art Unsustainable

As a form of outdoor public art, murals are constantly exposed to the natural environment and are highly susceptible to external factors such as erosion from wind and rain, weathering from sunlight, accidental damage, and staining. At the same time, they face the risk of demolition as part of urban and rural renewal projects; their sustainable preservation requires the support of a comprehensive, long-term maintenance mechanism (Liu Chang, 2023). However, in the vast majority of regions today, mural projects suffer from a tendency to "prioritize construction over maintenance." While ample human and material resources are invested during the construction phase, there are no mechanisms for subsequent maintenance, repair, or renewal once the project is completed. The long-term lack of professional maintenance has led to problems such as fading images, peeling walls, damaged designs, and stain coverage in a large number of high-quality murals. As a result, the complete cultural narratives depicted in these works are gradually becoming fragmented and blurred, and their function as carriers of memory continues to weaken. At the same time, in spatial renewal projects such as the renovation of older urban neighborhoods, the enhancement of rural landscapes, and the construction of road and utility networks, a large number of distinctive murals that embody local collective memory lack specific protection regulations. As a result, they are arbitrarily demolished, covered up, or replaced, causing the irreversible loss of regional non-verbal cultural memory and leading to a rupture in the transmission of cultural context and emotional continuity embodied by these murals. Furthermore, the absence of a dynamic update mechanism has led to the long-term stagnation of mural content, rendering it unable to adapt to the demands of the times and the evolving needs of public memory, and causing it to gradually lose its contemporary vitality.

5.4 Creation Disconnected from Public Opinion; Insufficient Alignment with Public Sentiment

Currently, most mural projects adopt a top-down creative model characterized by "designer-led, government-driven, and public-excluded" approaches. The entire creative process lacks the participation and input of local residents, resulting in a severe disconnect between mural content and the public's authentic memories and emotional needs (Wang Xu, 2024). Creative teams often rely on their own professional aesthetic judgment, design experience, and general creative logic to produce their work, without conducting research visits, focus groups, or public opinion surveys to uncover local residents' unique childhood memories, folk cultural perceptions, and emotional needs. Nostalgic memories vary significantly among residents of different regions and age groups: the elderly tend to focus on traditional farming and life in old alleys; middle-aged and young adults focus on childhood memories of their hometowns and folk festivals; and adolescents focus on new rural trends and distinctive cultural perceptions. Standardized and subjective creative models cannot accommodate the diverse memory needs of different groups. As a result, the final mural content becomes detached from the public's real-life experiences, making it difficult to evoke widespread emotional resonance. This leads to low public acceptance and a lack of identification, fundamentally limiting the full realization of murals' value as non-verbal carriers of emotion and memory.

6. Optimization Pathways for the High-Quality Development of Murals as Non-Verbal Carriers

6.1 Focus on Core Content to Create High-Quality Cultural Vessels

To overcome the dilemma of homogenization and elevate murals from "visual decoration" to "cultural vessels," the key lies in adhering to the principle of localized creation and constructing a deep narrative system. In the early stages of creation, a systematic mechanism for cultural tracing and research must be established. Creators should delve into villages, towns, and communities to comprehensively document historical developments, folk customs, distinctive industries, scenes of daily life, and symbols of collective memory, thereby building a repository of local cultural resources and eliminating the use of cookie-cutter templates. Based on local resources and in conjunction with spatial characteristics and community structures, customized creations should be developed. In rural areas, the focus should be on exploring local elements such as agriculture, folk customs, ancient villages, and intangible cultural heritage; urban communities should focus on contemporary memories, everyday life, and neighborhood culture; and cultural tourism areas should integrate landmark landscapes and distinctive industries into their creative work.

At the same time, creators should shift their mindset, placing cultural heritage and emotional expression at the core, and ensuring that the exploration of deeper meanings permeates the entire creative process. They must

thoroughly trace the historical origins, folkloric symbolism, and emotional significance of local visual elements to ensure that every visual detail is culturally grounded. Through coherent visual compositions, progressive scene design, and a unified thematic core, they should tell the region's story in its entirety, forming a non-verbal cultural narrative chain. While upholding the traditional essence, they should appropriately incorporate modern aesthetics and contemporary concepts to achieve an organic fusion of tradition and modernity. This approach not only evokes nostalgic memories but also aligns with contemporary aesthetic demands, enabling the murals to accurately convey the region's unique collective memory and cultural spirit.

6.2 Strengthen Operation and Maintenance to Build a Sustainable Heritage Ecosystem

To ensure the sustainable development and continued vitality of murals as non-textual cultural carriers, it is necessary to establish a long-term, full-process mechanism that combines "institutional safeguards" with "public participation." On the one hand, a dedicated protection system should be established to include high-quality murals that embody core collective memories in the public cultural heritage protection registry, clearly defining protection standards and responsible entities. A professional operations and maintenance team should be formed, dedicated funding allocated, and mechanisms for regular inspections, repairs, and cleaning established to ensure the integrity and aesthetic appeal of the murals. At the same time, a dynamic update mechanism should be established to periodically renovate and supplement aging murals in line with urban and rural development and contemporary aesthetic trends, incorporating the achievements of the new era. On the other hand, a bottom-up, public participation mechanism should be established, centered on the authentic memories and emotional needs of the public. Break away from the model of sole designer dominance by incorporating public opinion surveys, public participation, and community reviews into the entire creative process. In the early stages, widely collect residents' nostalgic memories and creative suggestions through forums, interviews, and questionnaires; during the middle stages, open channels for proposal review and invite resident representatives to participate in vetting and optimization; simultaneously, organize public art creation activities to encourage residents to participate in hand-painting, thereby strengthening emotional connections. In the later stages, a public feedback mechanism is established to continuously optimize the creative process and ongoing maintenance. Through public participation, mural art is truly rooted in people's daily lives, maximizing its core values of emotional comfort, memory consolidation, and community identity, thereby achieving the sustained preservation and innovative development of local cultural memory.

7. Conclusion

As a living, non-verbal medium rooted in urban and rural public spaces and closely tied to people's daily lives, mural art builds a warm bridge between contemporary people and traditional memories, local culture, and nostalgia for one's hometown, allowing fading rural memories to be preserved, scarce local cultural contexts to be sustained, and widespread social nostalgia to find a home. Against the backdrop of continuous innovation in public art and the ongoing deepening of urban and rural cultural development, mural art will continue to demonstrate its multidimensional value as a medium, serving as a vital tool for revitalizing urban and rural culture, fostering community bonds, and preserving regional cultural heritage. It will provide enduring support and assistance for the development of high-quality, people-centered urban and rural environments, the living transmission of local culture, and the governance of social emotions.

References

Brown D, Davis J. (2019). Cultural Memory and Nostalgia Carriers in Urban Renewal. *Journal of Urban Cultural Studies, 6*(2), 45-62.

Chen Li, Liu Yang. (2024). Contemporary Renewal and Spatial Revitalization of Rural Nostalgic Murals. *Rural Development Research, 8*(2), 36–43.

Fan Ze. (2024). Conceptual Shortcomings and Optimization Strategies in Contemporary Nostalgic Mural Creation. *Art Observation, 9*(1), 58–63.

Guo Xin. (2021). Visual Intuition and Emotional Communication in Public Art. *Visual Arts Research, 10*(3), 29–35.

Han Min. (2021). Collective Memory and the Construction of Community Cultural Identity. *Sociological Studies, 35*(6), 89-96.

Huang Bo. (2022). Situational Immersion and Nostalgic Emotional Resonance in Spatial Visual Art. *Research on Culture and Art, 11*(2), 44-50.

Jiang Yu. (2024). Pathways to Revitalizing Urban and Rural "Gray Spaces" Through Mural Art. *Urban Planning Review, 48*(3), 77-83.

Johnson S. (2018). The Evolution and Cultural Connotation of Nostalgia Theory. *Journal of Social Psychology, 34*(5), 312-320.

Li Yue, Wang Qiang. (2022). The Emotional Redemptive Function of Visual Nostalgia Media. *Journal of Emotional Psychology, 10*(4), 55-61.

Liu Chang. (2023). Developmental Challenges and Countermeasures for Contemporary Rural Mural Art. *Research on Art Education, 18*(5), 39-44.

Liu Hao. (2024). The Value of Nostalgic Murals in Preserving Local Culture. *Research on Traditional Culture, 12*(1), 76-82.

Miller T. (2020). Public Murals and Community Memory Construction. *International Journal of Public Art, 8*(2), 102-110.

Smith L. (2017). Narrative Memory and Community Identity of Urban Murals. *Urban Art Research, 5*(3), 67-75.

Wang Shuo. (2023). Public Art and the Consolidation of Regional Collective Memory. *Regional Culture Studies, 11*(3), 59-65.

Wang Xu. (2024). Deficiencies in Demand-Oriented Approaches to Contemporary Nostalgic Public Art Creation. *Art Review, 7*(2), 49-54.

Wu Jie. (2023). The Medium-Specific Attributes and Emotional Value of Nostalgic Murals. *Contemporary Art Studies, 12*(4), 66-72.

Xu Peng. (2023). The Problem of Homogenization in Rural Murals and Pathways to Differentiated Creation. *Landscape Design Research, 15*(3), 44-49.

Yang Lin. (2020). Nostalgic Anxiety and Spiritual Solace in the Process of Urbanization. *Review of Social Psychology, 8*(4), 77-83.

Zhang Yuan. (2022). A Study on the Continuity of Nostalgic Culture Across Urban and Rural Contexts. *Urban and Rural Planning, 46*(5), 99-105.

Zhao Lei. (2023). The Emotional Warmth and Artistic Expression of Nostalgic Murals. *Research on Art Creation, 9*(3), 62-67.

Zhou Yang. (2023). The Value of Public Mural Art in Spatial Revitalization. *Urban Space Design, 12*(2), 44-50.

Paradigm Academic Press
Art and Society
ISSN 2709-9830
JUN. 2026 VOL.5, NO.2

Visualizing Familial Order: Female Donor Images in Mogao Cave 156 at Dunhuang

Junrui Zou[1]

[1] Xi'an Academy of Fine Arts, Xi'an, Shaanxi 710000, China

Correspondence: Junrui Zou, Xi'an Academy of Fine Arts, Xi'an, Shaanxi 710000, China.

doi:10.63593/AS.2709-9830.2026.06.003

Abstract

Mogao Cave 156 preserves a remarkable group of clearly identified female donor images, providing important evidence for understanding kinship, social identity, and visual culture in Dunhuang during the Guiyijun (Return-to-Allegiance Army) period. Previous studies have primarily examined Cave 156 through the perspectives of political history, Zhang Yichao's military authority, donor identification, and patronage systems, while the role of female donor imagery in constructing lineage identity and family memory has received comparatively limited attention. Focusing on the female donors in Cave 156, this article investigates how inscriptions, cartouches, iconographic features, and spatial organization contributed to the formation of a lineage-oriented visual program within the Buddhist cave space. It argues that these female images functioned not merely as representations of individual patrons but as visual instruments through which the Zhang clan articulated kinship relations, political legitimacy, and collective memory. Their integration into the cave's broader representational structure reveals that female donors occupied an active role in the construction of family identity, expanding beyond the subordinate visual positions often assigned to female donors in earlier grotto contexts. Through an analysis of female donor imagery, this study demonstrates that Cave 156 reflects the changing function of donor representation during the Guiyijun period, when Buddhist cave construction increasingly became a medium for expressing family order, lineage continuity, and social authority. The female donors of Cave 156 thus provide a significant perspective for examining the intersection of gender, family, and political identity in Dunhuang visual culture.

Keywords: Mogao Cave 156, female donors, lineage identity, visual culture, Dunhuang

1. Introduction

In 848 CE, amidst internal strife within the Tibetan Empire, Zhang Yichao, a native of Shazhou, led a popular uprising that overthrew seventy years of Tibetan rule and reclaimed the eleven prefectures of Hexi. Three years later, the Tang court established the Guiyijun (Return-to-Allegiance Army) at Shazhou, appointing Zhang Yichao as the Military Commissioner (Jiedushi) over the region, with his seat of administration in Dunhuang. With the rise of the Guiyijun regime, the Mogao Grottoes entered a new golden age. From Zhang Yichao's uprising in 848 until the collapse of the Guiyijun regime following the Western Xia conquest of Shazhou in 1036, a total of 71 new caves were excavated and 11 earlier caves were renovated at Mogao. Among these, Mogao Cave 156 stands out as one of the most representative monuments of this era. It was commissioned by Zhang Huaishen—nephew of the first Military Commissioner—to commemorate his uncle's historic achievement in recovering Hexi. Cave 156 is rich in content, mixing Exoteric and Esoteric elements, with its front chamber, corridor, and main chamber well preserved, displaying diverse artistic styles and exquisite workmanship. What attracts the most attention is the large number of donor images painted on the walls of the corridor and the main chamber. Compared with caves of previous dynasties, the number and scale of female donors in Cave 156 significantly increased, occupying important spatial positions in the cave and becoming an important part of the visual space of the cave. From the perspectives of spatial research and iconographic research, this paper discusses the identity composition, spatial

layout, and cultural significance of the female donors in Cave 156, and analyzes how family relations in real society were transformed into visual order in Buddhist space.

2. Cave 156 and the Transformation of Donor Imagery into Lineage Representation

After the establishment of the Guiyijun regime, Dunhuang moved from the Tibetan period to the rule of the Zhang kinship, local society restored stability, and economic activities re-developed. This transition restored local stability and revitalized economic activities; yet, the enduring and profound impact of Tibetan Buddhist policies on local society and livelihoods remained undeniable. Consequently, Guiyijun generals and local gentry clans commissioned numerous "merit caves" (gongde ku) to pray for blessings and commemorate their achievements. Concurrently, as the de facto rulers of the Guiyijun regime, the Zhang kinship urgently needed to consolidate their political legitimacy and social influence through diverse channels. The construction of Cave 156 was completed precisely under this background.

During the construction of merit caves, in order to accumulate merit, seek protection, display family prestige, and leave names for future generations, donor images gradually became one of the core contents of the caves. Donors painted images of themselves and their families inside the caves, symbolizing that they were guarding the Buddhist Dharma at all times and worshiping the Buddha day and night, thereby praying for the Buddha to bless the family's peace. The construction of large caves required a large amount of funds and was usually presided over by great and prominent families. By rendering donor portraits alongside detailed cartouches of names and official titles on the murals, these caves became potent symbols for flaunting lineage pedigree and projecting local political and economic dominance. Thus, the merit cave served not only as a physical vehicle for devotees to manifest their pious faith but also as a medium through which they hoped to preserve their visages for eternity, achieving a form of "spiritual immortality."

Cave 156 contains a total of 37 male and female donor portraits, constructing a clear and comprehensive family genealogy. The emergence of this vast corpus of donor imagery demonstrates, on one hand, that members of the Zhang kinship were devout Buddhists who commissioned grottoes to accumulate merit and seek spiritual liberation in the afterlife. On the other hand, it reflects how the Zhang kinship embraced the pervasive Buddhist fervor of the era as a contemporary trend, while simultaneously projecting Zhang Yichao's monumental achievement in reclaiming the Hexi region. Crucially departing from preceding dynasties, the female donor images within this cave are no longer confined to marginalized corners; instead, they occupy primary spatial zones, such as the main chamber and the corridor, alongside their male counterparts. Functionally, these female donor portraits in Cave 156 served not only a religious, devotional purpose but also operated as mechanisms for chronicling clan history, broadcasting lineage status, and reinforcing collective family memory.

3. Female Donors and the Construction of Zhang Clan Genealogy

Based on extant cartouches and iconographic evidence, the female donors in Cave 156 comprise Zhang Yichao's mother, wives and concubines, sisters, nieces, and daughters, thereby encompassing pivotal female figures across multiple generations of the Zhang clan. In contrast to the limited presence of female donors found in standard caves, Cave 156 presents a far more comprehensive composition of individual identities, functioning effectively as a visual manifestation of a family genealogy.

On the north wall of the corridor, three standing female donor portraits are arranged facing the main chamber. The blue-ground small cartouche for the first donor has been blackened by smoke and lost; however, within a stele-shaped cartouche frame complete with a canopy and a base, the inscription reads: "Imperially Decreed Grand Lady of Song State, Mistress of Henei Commandery, Lady Song of Guangping, whole-heartedly makes offerings" (Chi Songguo Henei Junjun Taifuren Guangping Songshi yixin gongyang). Because the format of this cartouche is virtually identical to that of the first donor portrait on the south wall of the corridor—which depicts Zhang Yichao himself—it is inferred that this figure represents Zhang Yichao's principal wife, the "Lady of Song State." The second donor features a base-less and canopy-less small cartouche on a dark grey ground, inscribed with: "Niece Taizhen, Eleventh Daughter, whole-heartedly makes offerings," identifying her as the daughter of Zhang Yichao's elder brother, Zhang Yitan. The cartouche for the third donor is no longer extant; based on her spatial positioning, she is deduced to be one of Zhang Yichao's own daughters, who was younger than his niece. Arranged in sequence, these three female figures constitute a visual display of different generations within the domestic lineage.

Beneath the niche on the west wall of the main chamber, the donor portraits extend sequentially toward the north and south sides, centering around an altar table and an incense burner. The south side features two Buddhist monks and three male donors, while the north side consists of Buddhist nuns and female donors; the monastic and secular realms are thus physically demarcated, standing on two distinct rugs. On the north side, the cartouche of the fourth donor portrait reads: "Dharma Master and Altar-Ascended Eminent Nun Liaokong, Elder Sister [of Zhang Yichao]" (Zishi dengtan dade ni falu Liaokong), confirming her identity as Zhang Yichao's sister who entered the Buddhist monastic order. Following the nun portraits, the third, fourth, and fifth female donor figures are

respectively inscribed with: “Lady Suo, whole-heartedly [makes offerings],” “Lady Fu, whole-heartedly makes offerings,” and “Lady [illegible], whole-heartedly makes offerings”—all of whom were wives of Zhang Yichao. Based on their spatial positioning and relationship to the male donors aligned on the south wall, it is inferred that the first two female donor figures represent Zhang Yichao’s other wives: Lady Song of Guangping and Lady Lu of Fanyang. Consequently, this section juxtaposes the female members of the Zhang clan who took monastic vows alongside his multiple wives.

Above the entrance on the east wall of the main chamber are four donor portraits. The south side features three male donors; according to their cartouches, they depict Zhang Yichao’s father (Zhang Qianyi), his elder brother (Zhang Yitan), and the cave patron himself (Zhang Yichao). The north side displays a single female donor, inscribed as: “The deceased mother, posthumously title Grand Lady of Song State, née Chen, whole-heartedly makes offerings” (Wangmu zeng Songguo Taifuren Chenshi yixin gongyang), accompanied by two attendants standing behind her. Given the identity of the cave owner, this “deceased mother, née Chen” is deduced to be Zhang Yichao’s mother.

Spanning the lower sections of the north wall and the northern part of the east wall in the main chamber is the Lady of Song State’s Procession, which bears the inscriptions: “The Procession of Lady Song, Mistress of Henei Commandery, Lady of Song State” (Songguo Henei Junfuren Songshi chuxing tu) and “The Carriages and Horses of the Lady of the Minister of Works, née Song” (Sikong furen Songshi xingli chema). Secular in essence, this grand and intricate composition departs from conventional static donor portraiture, mirroring the contemporary ritual systems and actual historical tableaus. Positioned directly opposite Zhang Yichao’s Military Procession, it underscores the exulted status of the Lady of Song State, thereby broadcasting the patron’s illustrious religious background and political preeminence. The female donor images in Cave 156 together constitute a relatively complete female family genealogy, with obvious characteristics of a clan-sponsored cave. The connection between religious space and family space has become increasingly close, and female family members have become an important part of the construction of the Zhang clan’s historical image.

4. Clothing, Cartouches, and the Visual Encoding of Female Identity

The facial features and clothing of the three female donor images in the corridor are relatively similar, and the primary-secondary relationship of the figures is distinguished only by detailed decorations. The first figure, Lady Song, wears a high-waisted skirt, a red large-sleeved jacket (ru), and a high coiffure topped with a floral crown; a grand canopy hovers above her head, and her face is adorned with facial patches in the Uighur style (Huihu zhuang). The second figure, clad in a large-sleeved jacket with small floral patterns on a blue ground and a high-waisted skirt, holds a flower tray in offering; her head and facial adornments mirror those of the first figure. The third figure is virtually identical to the second, though the coloration of her garments appears grayish-black, likely an artifact of flaking and pigment discoloration. The female donor images beneath the niche on the west wall of the main chamber exhibit an even more pronounced degree of schematization. The first four bhiksunis (nuns) possess nearly identical vestments and facial countenances. They are all tonsured, and only the offering objects held in their hands are different: the first one holds a long-handled incense burner, and the second one holds a flower tray for offering. The clothing folds are all similar; they wear high-waisted skirts inside and kasaya (robes) outside, distinguished by the color of the clothing. On the second rug, the five secular figures display near-identical facial countenances. With high coiffures adorned with flowers, they are attired in luxurious high-waisted silk skirts and large-sleeved shirts embellished with diverse patterns; the textiles appear soft, noble, and magnificent as the figures reverently proffer fresh blossoms. Above the entrance on the east wall, the portrait of Zhang Yichao’s mother shares a fundamentally similar sartorial style with the figures beneath the west niche, featuring a large-sleeved jacket, a high-waisted skirt, a translucent silk shawl, and cloud-tipped shoes; her hair is styled in a high coiffure as she holds an incense burner with both hands in pious offering.

Judging from the donor portraiture in Cave 156, the visages of the figures exhibit a distinct tendency toward stylistic conventionalization; there is minimal differentiation in the rendering of facial features among individual female donors. Conversely, pronounced variations manifest in their attire, devotional objects, and spatial positioning. This phenomenon demonstrates that the artisans’ focus was directed not toward individual physiognomy, but rather toward institutional identity. Consequently, the cartouches, garments, and spatial placement coalesced into a unified and critical framework for identifying the respective figures.

The change of concepts reflects the change in the function of donor portraits. When the cave assumed the function of family commemoration, the most important value of the figures did not lie in reproducing their true appearance, but in accurately expressing their position in the family. The hierarchy of clothing, the norms of posture, and the arrangement of figures together formed a complete system of identity expression, allowing viewers to recognize the social and family relations in which the images were situated. The iconographic shaping of female donors in Cave 156, on the one hand, preserves important visual materials of Dunhuang female images during the Guiyijun period; on the other hand, it is also an important part of the visualization of family order.

5. Lady Song's Procession and the Political Visibility of Women

Among the female donors in Cave 156, the status of Zhang Yichao's wife, Lady Song, is particularly important. She not only appears multiple times in the donor procession, but also has an independent Procession Chart of Lady Song, Henei Commandery Lady of Song State, which has a grand scene, numerous figures, an innovative composition, and vivid shapes. The overall picture is divided into three main parts: starting from the west side of the north wall, the front part consists of pole-balancing acrobats, followed by a dance and music troupe, including four dancers and eleven musicians accompanying them; Behind the troupe advances a cohort of equestrians and attendants, trailed by baggage carts and horses. In the central section, Lady Song rides on horseback; adhering to the traditional artistic convention of hierarchical proportion (where major figures are rendered larger than minor ones), she occupies the absolute visual core of the composition. Depicted with an exalted stature, Lady Song wears a floral crown adorned with hairpins, a cyan large-sleeved robe with a cross-collar, and a high-waisted skirt tied at the chest; she sits atop a robust, long-maned white horse. The line work defining her drapery folds, headdress, and the horse's mane remains sharp and crisp, followed by a rearguard of armed escorts. The final section transitions to the northern side of the east wall, where six mounted retainers lead the transport baggage, and the entire procession culminates in a scene of mounted soldiers practicing archery. As scholarship notes, "utilizing procession scenes to flaunt one's achievements, power, and status was the primary motivation for Guiyijun rulers to commission this specific iconographic schema."

Grand in scale, the Lady Song's Procession encapsulates carriages, horses, guards of honor, attendants, and entertainment performances, presenting a solemn and majestic processional tableau. The composition skillfully balances density and openness in its grouping of figures, marking a culmination of traditional Chinese equestrian and figure painting. Positioned directly opposite Zhang Yichao's Military Procession, both murals adopt a horizontal scroll format, each measuring an impressive 8.2 meters in length and 1.05 meters in height; their sheer human scale, the length of the entourages, and the opulence of decorative details collectively project the family's political authority and social ascendancy.

The true significance of the Lady Song's Procession, however, lies in its deliberate integration of a core female family member into the expressive framework of political power. By appearing side by side with Zhang Yichao, Lady Song became a part of the overall authority of the Zhang kinship, reflecting the change of family concepts during the Guiyijun period. Although women were still within the traditional family structure, their iconographic value had exceeded the family subordinate role in the general sense, and they became important media for maintaining family memory and strengthening family legitimacy.

6. Spatial Order and the Visual Construction of Kinship

One of the most noteworthy features of Cave 156 lies in the fact that its donor layout roughly forms a fixed pattern, painting numerous figures from the cave merit patron's family on each wall surface, where males are located on the south side and females are on the north side. The reorganization of family relations in religious space transforms the clan order in real society into a visual order in Buddhist space, and the two form a relatively stable corresponding relationship.

In the cave, the social relations of the donors are expressed through spatial positions, and female family members are arranged in different positions in the cave according to their identity and status, reflecting real kinship relations. Thanks to this, viewers entering the cave can understand the internal relational structure of the Zhang kinship through the arrangement of figures. The corridor, the main chamber, and the area beneath the niche collectively serve as venues for displaying lineage members, synthesizing into a visual space of a deeply commemorative nature.

Observing the distributional patterns of these female donor portraits reveals that the clan members were integrated into a cohesive whole characterized by explicit hierarchies and structure; the Buddhist space thus operated as an essential medium for broadcasting the family framework. From this perspective, family members attained a form of perpetual presence through their iconographic surrogates, while the cave transformed into a critical locus for preserving and transmitting familial memory—ultimately achieving the visual manifestation of a family genealogy.

7. Conclusion

In conclusion, the female donor iconography in Cave 156 offers a fresh vantage point for investigating the social fabric and familial culture of Dunhuang during the Guiyijun period. The identities of female donors in the cave are reflected through cartouche inscriptions, iconographic positions, and spatial layout, transforming real-world family member relations into visual expressions in the grotto space. Drawing on theories of visual culture and collective memory, this article approaches Cave 156 not simply as a Buddhist devotional space but as a site where lineage identity was materially and visually constructed. The images of female donors further reflect the multiple meanings of religious and political power in the construction of family order.

The female donor images in Cave 156 embody the family-centered mode of cave construction. The identity markers, arrangement relationships, and ritualistic expressions of the figures together constitute the image presentation of the Zhang kinship in the Buddhist space. Through these images, it can be observed that while the Dunhuang grottoes during the Guiyijun period undertook religious functions, they also became important spaces for constructing the expression of family order and confirming social identity. In this process, the female images connected the individual, the family, and the local social relations, showing the side of grotto art participating in the construction of social order.

References

Books / Monographs

Dunhuang Academy. (1986). *Donor inscriptions of Dunhuang caves*. Cultural Relics Press.

Pelliot, P. (2007). *Notes on the Dunhuang caves* (S. Geng, Trans.). Gansu People's Publishing House.

Rong, X. (2015). *A study of the Guiyijun regime: Historical investigations of Dunhuang during the Tang and Song periods*. Shanghai Ancient Books Publishing House.

Tan, C. (1993). *Marriage culture in Dunhuang*. Gansu People's Publishing House.

Wong, D. C. (2018). *Buddhist pilgrim-monks as agents of cultural and artistic transmission: The international Buddhist art style in East Asia, 7th–10th century*. NUS Press.

Journal Articles

Li, G., & Sha, W. (2017). A reconsideration of the construction history of Mogao Cave 156. *Dunhuang Research*, (5), 49–56.

Sha, W. (2003). A study of donor images in Dunhuang caves during the Tibetan occupation period. *China Tibetology*, (2), 80–93.

Sha, W. (2007). Donor images and secular Buddhism in Dunhuang during the Tang and Song dynasties. *Dunhuang Research*, (4), 72–79.

Zhang, P. (2008). The activities of lay communities in the construction of Mogao Caves during the Tang and Song periods: Centered on donor images and inscriptions. *Journal of Dunhuang Studies*, (4), 123–131.

Zhang, X. (2008). The development and transformation of donor images in Mogao Caves: A study from the perspective of Buddhist history. *Journal of Dunhuang Studies*, (4), 93–103.

Zheng, B., & Xu, X. (2003). A study of polygamy among the military governors of the Guiyijun regime in Dunhuang during the late Tang and Five Dynasties. *Journal of Northwest Second University for Nationalities (Philosophy and Social Sciences Edition)*, (4), 38–45.

Copyrights

Paradigm Academic Press
Art and Society
ISSN 2709-9830
JUN. 2026 VOL.5, NO.2

The Triple Interactive Logic of Interactive Film and Television

Shulin Li[1]

[1] Northwest Normal University, Gansu, China

Correspondence: Shulin Li, Northwest Normal University, Gansu, China.

doi:10.63593/AS.2709-9830.2026.06.004

Abstract

As an emerging form of film and television art, interactive film and television breaks through the one-way communication mode of traditional works, forming three layers of interactive logic. The first layer, the logic of experiential interaction, manifests as the subject's selection of and participation in the object, as well as the object's feedback and incentive to the subject. The second layer, the logic of creative interaction, manifests as the creators' "dialogic" primary creation, the audience's "participatory" secondary creation, and the community's "discursive" tertiary creation. The third layer, the logic of aesthetic interaction, manifests in the aesthetic interaction of "a thousand faces," the aesthetic interaction of "repetitive dynamics," and the aesthetic interaction of "subject empathy." These three layers of interactive logic together constitute a completely new signifying form for interactive film and television art, expanding the narrative methods and expressive space of film and television art.

Keywords: interactive film and television, interactive film game, film and television experience, film and television creation, film and television aesthetics

1. Introduction

In December 2023, a Chinese interactive film and television work titled *Love Is All Around* sold 1.9 million copies and generated 80 million yuan in revenue on Steam, ranking second on that year's domestic game sales chart. This commercial success brought interactive film and television back into the spotlight. As early as 2012, the interactive film-game *The Walking Dead*, developed and published by Telltale Games, caused considerable stir in the industry. The release of *Life Is Strange* in 2015 prompted greater attention to the artistic value of such works. In 2018, *Detroit: Become Human*, developed by Quantic Dream, and Netflix's new interactive drama *Black Mirror: Bandersnatch* were launched, truly ushering interactive film and television into the public eye.

Today, what we refer to as interactive film and television comprises two main forms. The first is the "interactive film," defined as "a video form that hands control over the plot's development to the audience (player), adopts a multi-branch textual structure, and creates an interactive viewing experience". (Zhang Dewei, 2020) This type of work is based on a modernized refinement of traditional film viewing methods and, in terms of experience, leans more toward traditional cinema that contains multiple narrative directions. The second is the "interactive film-game," a genre that, "compared with the interactive film, embodies a fuller integration of 'film' and 'game'" (Chen Baijun & Huang Xinyuan, 2024), its experience mode retains the viewing sensation of a traditional film while emphasizing the incorporation of the strengths of games. Fundamentally, these two forms can be seen as two different paths in the production philosophy of interactive film and television, and the interactive film and television explored in this article encompasses both types. Today, with the development of interactive film and television, the boundary between games and films is gradually blurring, and its market acceptance is continuously expanding.

2. "Experiential Interaction": Gamified Film and Television Experience

Compared with traditional film and television, the interactivity of interactive film and television is most

intuitively reflected in its mode of experience. Unlike the one-way reception of the traditional big screen, interactive film and television, in terms of experience, places greater emphasis on the audience's (player's) active participation, active immersion, and active comprehension. In terms of current experience, the two-way interaction between subject and object in interactive film and television is primarily manifested in the subject's ability to select and participate with respect to the object, and on this basis, a "role-playing" experience different from that of traditional cinema is formed. At the same time, the object exerts incentive and feedback effects on the subject, using content such as "moral dilemmas" and "reward mechanisms" to provide rational and emotional regulation and guidance to the subject. The "two-way interaction" of these two mechanisms constitutes the main content of "experiential interaction" in interactive film and television.

2.1 The Subject's Selection of and Participation in the Object

"Selection" is the core of the experiential process of interactive film and television works. "Selection" refers to the audience's (player's) right, during the experience of interactive film and television, to actively choose specific content such as characters, the characters' ways of handling matters, and the direction of the plot's development. For example, in the work *Detroit: Become Human*, the audience (player) can freely choose among multiple characters to experience the narrative from different standpoints and perspectives. At the same time, within the narrative line of a single character, the audience (player) can choose different modes of behavior. Taking the character Connor as an example, as an advanced android, the audience (player) can freely choose whether to obey human commands without limit, even if those commands may deviate from their original intentions or restrict their freedom; or the audience (player) can decisively choose to pursue freedom, though this may entail greater risks. The creators do not provide the audience (player) with any hints during the selection process, and there is no single absolutely correct choice. The audience (player) must, based on their own judgment, continually push the narrative forward, ultimately arriving at an ending of their own creation. Regardless of the outcome, the audience (player) ultimately knows that it was the result of their own choices. This mode of experience not only endows the audience (player) with tremendous experiential freedom but also grants them the right to control the fates of the characters in the film and television work. It not only greatly enhances the audience's (player's) sense of immersion and engrossment, but also endows film and television works with greater possibilities for interpretation.

"Participation" is the key to the experiential process of interactive film and television works. Many such works allow the audience (player) to participate in the concrete actions of the story's development. Take *The Walking Dead* as an example: in the first chapter, when the audience's (player's) vehicle overturns, the work requires the audience (player) to use the controller to smash the window, thereby enabling the character to escape the overturned vehicle; after the character escapes, the audience (player) is required to use the controller to pick up a branch as a weapon. This form of "participation" strengthens the audience's (player's) identification with the characters in the drama, "creating an interactive 'role-playing' environment that allows the audience (player) to experience a sense of reality within a physically engaging interactive environment" (Wenqi Liu, 2023). This experiential form also further blurs the boundary between games and film and television, making it an "interactive film-game" in the practical sense. Beyond "interactive participation" in action, another mode of participation in interactive film and television is dialogic participation. As the primary means of advancing a film and television work's narrative, lines of dialogue are traditionally set in advance by the scriptwriter and director during the creative phase and then delivered by the actors—a process in which the audience (player) effectively has no participation. In interactive film and television, however, creators often provide not just one mode of dialogue but multiple dialogue options corresponding to different personalities. Again taking the first chapter of *The Walking Dead* as an example, when the audience (player) is in the police car at the very beginning of the experience, the police officer continuously converses with the audience (player), who can then select from numerous dialogue options the one that best fits their own personality; the film and television work, in turn, will react differently depending on the dialogue choices made by the audience (player). This experiential mode enhances the sense of immersion in the work and expands the personality dimensions of the protagonist, so that the characters' personalities in the work are no longer fixed. The audience (player) is no longer merely watching the characters in a story but is genuinely playing themselves.

The deepening of the audience's (player's) "selection and participation" in the work enables the realization of "role-playing" in interactive film and television works. In the traditional film and television experience, the audience (player) cannot make any change to the characters in the work in any objective sense and can only passively follow the development of the plot. This mode of experience hinders the audience's (player's) "empathy" toward the film's characters. When watching a film or television work, the audience (player) often brings their own "horizon of expectations," which inclines them more readily to empathize with plot elements that align with their past experiences, preferences, and judgments of good and evil. Conversely, if the characters or plot in the work conflict with the audience's (player's) prior expectations, this may lead to a negative appraisal of the film and thus make empathy difficult. In interactive film and television, however, the

determinative power of this horizon of expectations is severely weakened, because the development of the plot and the characters' judgments are entirely placed in the hands of the audience (player) to decide on their own. The audience (player) seems to be experiencing their own life, thereby finding it easier to generate empathy and forming a "role-playing" type of experience. In the interactive film and television work *You Are My Courage*, the audience (player) can immersively play Bear Grylls in an expedition and can subjectively choose the routes, methods, and decisions. Bear makes during the adventure, fundamentally influencing the entire developmental trajectory of the story. In this sense, the audience (player) is not experiencing a story but rather creating a story. Furthermore, the realization of "role-playing" stems from the popularization of personalized viewing devices. The empathetic effect in traditional film and television originates from the audience's (player's) self-mirroring of the encounters or actions of the characters in the film. This self-mirroring arises more from the audience's (player's) psychological processes—the audience (player) expects and endorses the choices made by the characters on screen. In interactive film and television, however, this empathetic effect originates from the audience's (player's) direct "role-playing." "In 1966, a Czech director made the first interactive film in history, *Kinoautomat*." (Ni Taile, Chen Yingshuang & Wu Jinlong, 2021) In this film, the director creatively used a red-and-blue selection device. Whenever the film reached a critical decision point, the live audience voted through the red-and-blue selector for the direction they most wanted to see, and this mode of selection constituted the earliest "role-playing" experience in interactive cinema. However, the disadvantage of this method was also obvious—no matter the outcome of the vote, the wishes of some portion of the audience would always be ignored. This format can in fact only be regarded as an early experiment in interactive cinema, an attempt at "group-customized plotlines" weighted toward the majority, and was still far from being worthy of the name interactive cinema. With the passage of time, the emergence of modern media such as home PCs, mobile phones, and tablets changed this situation. Compared with the large-scale screenings in cinemas in the early days, modern media are more suited to application in personal spaces. This personal application directly gave rise to the "role-playing" found in today's interactive cinema. Today, the experiential mode of interactive cinema mainly relies on "direct touch" or "external controllers" as the primary operating methods; what is more, with the emergence of all-in-one VR operating devices, eye-tracking control, voice control, and various other modern operating methods continue to appear. The audience (player) no longer needs to sit in a cinema; rather, they engage in personalized film and television appreciation in a private space through personal devices. In this process, their preferences and choices are entirely pure, personalized, and customized, free from the influence of others, which strengthens individualized expression. The audience's (player's) mode of experience is no longer constrained by the outside world but turns to focus purely on the inner self, and the film and television experience shifts from passive to active, thereby finally presenting a complete "role-playing" experience.

2.2 The Object's Incentive and Feedback to the Subject

In interactive film and television, the object's feedback to the subject can be divided into intrinsic feedback and extrinsic feedback. Intrinsic feedback typically acts upon the player's inner world, appealing to the player's emotions and provoking moral reflection and decision-making. Extrinsic feedback, on the other hand, is directly manifested in the form of concrete content following any impact the player has on the filmic world, often consisting primarily of rewards such as items, props, and plot developments.

The "moral dilemma" is one of the primary forms of intrinsic feedback that interactive film and television works provide to the subject. Ever since the art of cinema gradually gained acceptance worldwide, the moral values presented by films have become a much-discussed topic. The moral values conveyed by traditional film and television are constrained by multiple factors, including the creator's personal views, the era, and politics, and these values are determined mainly by factors other than the audience (player). The audience (player) serves merely as a passive recipient, unable to actually influence the values conveyed by the film. By contrast, the emergence of interactive film and television renders this situation no longer fixed. In interactive film and television, every morally charged decision is made and chosen by the audience (player) themselves, and the resulting consequences are likewise tied to the audience's (player's) individual choices. Take the work *Life Is Strange* as an example: at the end, the audience (player) must make a difficult choice between sacrificing the town or risking everything to save it. This kind of "trolley problem" dilemma has become one of the major attractions of interactive film and television. Every decision made by the audience (player) will be subjected to the test of their own conscience.

The introduction of a "reward mechanism" expands the object's extrinsic incentive methods toward the subject in interactive film and television works, greatly enhancing the sense of immersion in the work. The "reward mechanism" originates from video games. In video games, corresponding reward mechanisms are necessary for every action the audience (player) takes in order to drive them to continuously explore proactively and to constantly achieve new accomplishments. During this process, the audience (player) is in an active state, continually seeking the "incentives" they need. In traditional film and television works, this reward mechanism is implicit, depending on whether the film satisfies the audience's (player's) "horizon of expectations"; the

audience (player) passively receives "incentives" during the viewing process, thereby being continuously drawn to the film and gradually empathizing with its plot. In interactive film and television works, the reward mechanism of video games is introduced. Taking *The Walking Dead* as an example, the work establishes an extremely horrifying and terrifying post-apocalyptic scenario from the very beginning. As the audience (player) explores continuously, they will acquire more supplies, befriend more companions, obtain better weapons, and gradually piece together the clues to the world's collapse. Throughout this process, the audience's (player's) emotions constantly fluctuate with the plot's development, and the numerous "incentives" prompt the audience (player) to keep exploring. The audience (player) feels as though they are actually in the apocalyptic world, rather than merely watching or experiencing a film and television work, thereby generating an immense "empathy" effect.

3. "Creative Interaction": The Multilateral Creative Process

Another important difference between interactive film and television and traditional film and television lies in their multilateral creative process. The creative process of interactive film and television has undergone considerable change compared to that of traditional film and television. Whereas traditional film and television production is typically initiated by investors or directors and follows a strongly "centralized" creative approach, the creation of interactive cinema more clearly exhibits a "decentralized" character, forming a tripartite mode of interactive creation involving the lead creators, the audience (players), and the community.

3.1 The Creators' "Dialogic" Primary Creation

When creating, the lead creators of interactive film and television often operate on the basis of multi-party "dialogue." First, unlike traditional film and television creators, who, during the creative process, tend to follow primarily the will of the director or screenwriter, the creators of interactive film and television must give greater consideration to the audience's (players') reception and reactions. The creators need to constantly engage in "dialogue" with an imagined audience (player), anticipating the audience's (player's) responses and choices, and based on these predictions, refine the corresponding plot lines. Multiple narrative lines intersect and are arranged together to form a vast narrative web—a narrative mode also known as "arborescent narrative."

The creative team behind *Detroit: Become Human* remarked in a subsequent interview: "This is a new way of thinking. Imagining that the audience (player) can do something in the story, having space to express themselves, is very interesting, and that is the most important point. Because what we want to do is not to have the audience (player) listen to a story, but to have the audience (player) be in the story—this is a very different thing." [David Cage: "After *Detroit: Become Human*, We Have New Insights into Interactive Storytelling," What's Happening in Mobile Gaming, May 5, 2021.] This reveals the key factor that creators must consider during creation—how to place the audience (player) within the story. This causes the entire film creation process to revolve around the audience (player), engaging in continuous dialogue with the imagined audience (player), thereby constantly optimizing the developmental direction and dynamics of the "arborescent structure." This creative approach extends the state of the audience's (player's) absence characteristic of traditional film creation, shifting cinema thoroughly from "creator-centered" to "audience-centered," constituting a monumental transformation in film and television creation.

3.2 The Audience's "Participatory" Secondary Creation

Interactive film and television elevate the "secondary creation" of the work to a position of greater importance. From the perspective of reception aesthetics, the production of an artwork cannot be separated from the reader's secondary creation, "for it is only through the process of the reader's mediation that the work enters a continuously changing horizon of experience." (H. R. Jauss, 1987) In traditional film and television, when the audience (player) aesthetically engages with a work, they undertake secondary creation in their mind based on their own understanding, thereby producing an aesthetic creation derived from and even transcending the work. Interactive film and television, however, materialize this "secondary creation" into reality: through concrete choices about the direction of the story's development, the audience (player) de facto determines the plot's course and the fates of the characters, fundamentally influencing the connotations that the story aims to express. For interactive film and television, a complete work must be handed over to the audience (player) for secondary creation before it can truly achieve completeness; otherwise, it is merely an enormous audiovisual repository loaded with countless images and texts. Only when the audience (player) actually participates in the story and strings together the "scattered," "arborescent" multitude of narrative threads is the interactive film and television work truly completed in the proper sense. Therefore, in interactive film and television, the audience's (player's) secondary creation directly determines whether the work is complete. The audience's (player's) secondary creation is elevated to a position of equal importance to that of primary creation.

"Repeated experience" is the focus of the audience's (player's) secondary creation. Although traditional film and television also possess the value of repeated experience—and one can likewise gain further aesthetic insights

through repeated viewing—ultimately, the repeated experience of traditional film and television is limited to reinterpreting the original text, and is in fact a secondary excavation of content "unnoticed" or "difficult to perceive." The repeated experience of interactive film and television, by contrast, becomes an integral and necessary part of experiencing the complete work itself. Because the multiple branching and intersecting narrative threads mean that the audience (player) cannot fully experience the entire story in a single playthrough, the audience (player) needs "multiple experiences" and "repeated experiences," making different choices and following different storylines, in order to see the entire story in its entirety. Take *Detroit: Become Human* as an example: at the very beginning of the film, the audience (player) must select, from multiple optional characters, the one they wish to play. These characters not only have completely different life experiences and ways of handling matters, but the choices they face and the developmental directions of their stories are also entirely different. Only when the audience (player) has fully experienced one character and then, carrying the memories of that experience, immerses themselves in a new character and a new story, viewing the story's progression from another perspective and engaging in aesthetic interaction from a different standpoint, can they ultimately realize the complete aesthetic interaction of the interactive film and television work. Thus, in interactive cinema, "repeated experience" changes from "optional" to a "necessary" requirement for a complete experience. This not only extends the value connotations of the artwork itself but also embeds the audience's (player's) secondary creation into the work itself.

3.3 The Community's "Discursive" Re-Creation

The community's "discursive" re-creation has already become one of the distinctive features of interactive cinema. With the global trend toward networking, community culture in recent years has become ever more deeply intertwined with film and television. In the realm of traditional film and television, the broader "film circle" culture centered mainly on major film and television websites and forums is still flourishing, while community culture is rapidly moving toward a trend of "specialization" and "miniaturization." "Interest-based clustering" (Tong Xiaoxu, 2023) often takes a particular culture or "IP" as its core—such as "Cthulhu Mythos," "ACGN," or "Marvel"—and, furthermore, small-scale communities based on specific film and television works are gradually taking shape. Unlike traditional film and television, the community culture of interactive cinema manifests a high degree of "interactivity" on the basis of this miniaturization. Taking *The Walking Dead* as an example, when the work first landed on platforms like Steam, the production team immediately established a presence on the platform. Because *The Walking Dead* was released in the form of multiple episodes and seasons, after the launch of the first episode, the production team actively communicated with the audience (players) and used the many ideas generated by the community as reference for the next installment. In addition to the interactive creation between the audience (players) and the production team, since current interactive film and television works often bear a very high degree of similarity to games, the "mod" (modification and expansion) culture that exists in traditional game communities is equally applicable to interactive film and television. Using *Detroit: Become Human* as an example, the audience (player) can download "mods" shared by other audience members (players) in the community and implant them into the film and television work by technical means, thereby modifying the work itself to better suit the preferences of certain audience segments. For instance, by replacing the model of a certain character, one can experience a "romance" with a different person, or by substituting a character model, one can prevent a beloved character from dying in the plot. This mode of modification and creation, entirely rooted among the audience (players) themselves and realized upon the work itself, constitutes a form of "interactive community creation" and achieves the "decentralization" of film and television creation in the digital cinema era. Consequently, "the traditional aura of the author is gone forever, authority is displaced, and decentralization becomes conspicuous." (Gao Yuan, 2022) This has become one of the important features of today's interactive film and television.

4. "Aesthetic Interaction": Personalized Narrative Presentation

"Aesthetic interaction" is the final link in the interactivity of interactive film and television. In today's digital age, interactive cinema has redefined the relationship between the audience (player) and the work. Generally speaking, this relationship is manifested in three aspects. First, from the perspective of the subject, the experience of interactive cinema tends more toward a purely personalized "a thousand faces" aesthetic interaction. Second, from the perspective of the object, interactive film and television strengthen the multiple aesthetic values of the object itself, forming a "repetitive dynamic" aesthetic interaction. Third, from the psychological level, interactive film and television enhance the subject's "empathy," thereby achieving a higher level of aesthetic self-interaction.

4.1 "A Thousand Faces" Aesthetic Interaction

The "a thousand faces" aesthetic interaction is the most obvious distinction between the aesthetics of interactive film and television and that of traditional cinema. This distinction is primarily reflected in two dimensions: the aesthetic activity itself and the psychological experience. First, in terms of the aesthetic activity itself, the mode

of aesthetic interaction in traditional film and television has mainly taken the form of cinema screenings and collective viewing. This experiential mode is closely tied to the attributes of traditional cinema itself; scholar Shao Meng terms this the "group effect," meaning that "an individual's viewing behavior and perception are influenced by the group" (Shao Meng, 2018). Thus, apart from its intrinsic aesthetic value, traditional film and television's "collective" experiential mode has also become part of its aesthetic experience. By contrast, the aesthetic experience of interactive cinema is a purer, more "individualized" aesthetic activity. The selection mechanism of interactive cinema dictates that it cannot be experienced by multiple people simultaneously, because everyone's aesthetic preferences, tastes, and moral standards differ. Thus, there is no single choice path that can completely satisfy everyone, and it is impossible for the audience (players) to achieve a "common denominator" through communication. Therefore, interactive cinema is evidently unsuited to collective multi-person viewing. In fact, at the very beginning of interactive cinema, The Automatic Movie was screened in large cinema auditoriums; the method used at the time was to give each audience (player) a red and blue selection button, allowing them to choose the direction of the plot according to their own preference. Ultimately, through vote counting, the film's plot would advance by majority rule (with the minority submitting to the majority). However, it is obvious that this method merely "mechanically" selects the story based on the preferences of the majority, while the experience of the minority is ignored. Therefore, this form of interactive cinema only truly emerged on a large scale today—with the flourishing development of the internet and the convergence of home PCs, mobile phones, and tablets forming a "post-cinema" era of "four-screen integration." Personalized film and television presentation devices have provided an opportunity for the large-scale production of interactive cinema, becoming one of the driving forces behind the transformation and development of film and television. Second, from the perspective of psychological experience, the aesthetic interaction of interactive cinema exhibits conspicuous "exclusivity" and "privacy." In his work *Three Essays on the Theory of Sexuality*, Freud argued that in childhood, there exists "the manifestation of curiosity and the 'voyeuristic' desire," which "originates from the 'theft impulse' towards sex." (Sigmund Freud, 2004) This kind of privatized impulse plays an equally crucial role in film and television aesthetics. In the aesthetic experience of traditional film and television, the audience (player) "gazes" voyeuristically at the characters on screen (Laura Mulvey, 1975), thereby satisfying their own desires and needs and generating psychological pleasure. This need is even more pronounced in interactive film and television. The main reason is that, compared with traditional film and television, the "sense of shame" in interactive film and television arises not only from the "voyeuristic gaze" at the screen but also from the audience's (player's) private inclinations in selecting the content of the work. For example, in the interactive film and television work *Love Is All Around*, there are multiple female leads of different types and styles, and during the experience, the audience (player) may choose to develop a romantic relationship with one of them. Because each audience member's (player's) romantic inclinations and preferences differ, and their desires regarding women also vary greatly, different audience members (players) will choose completely different romantic trajectories and story development paths. And all of this is unknown and unwished to be known by others. "Romance," as a deeply intense private emotion that requires personalized "empathic" projection onto film and television characters, is merely one type among the various personalized needs found in numerous interactive film and television works. From this, one can see interactive film and television's demand for "privatized" scenarios, which makes it difficult to realize in a collective viewing environment. This private, personalized aesthetic interaction has thus become a significant feature distinguishing interactive film and television from traditional cinema.

4.2 "Repetitive Dynamics" Aesthetic Interaction

The complete experience of interactive film and television is constantly repeated and dynamically changing; therefore, its aesthetic interaction is also presented in the form of "repetitive dynamics." "Repetitive dynamics" is embodied in the aesthetic activity of interactive film and television. On the one hand, due to the existence of the "arborescent narrative," a single interactive film and television work often has multiple endings and multiple narrative threads, and each narrative line appears in an intersecting form. This means that the audience (player) cannot see the full picture of the story during the first experience. In fact, current interactive film and television works often exploit this to guide the audience (player) to experience the work multiple times, discovering content hidden behind the story lines from different angles. In *Detroit: Become Human*, the audience (player) can only see a portion of the story within a single story line; the work is divided into the "Markus line," the "Connor line," and the "Kara line," each of which possesses entirely different threads of story development. When the audience (player) subsequently starts a "second playthrough" and joins the story as another character, everything becomes clear. Such script designs compel the audience (player) who wishes to experience the complete plot to undergo "multiple playthroughs" and repeated experiences. Unlike traditional film and television, because the experiential mode of interactive film and television is primarily based on selection and interaction, multiple viewings will not feel as tedious as they do with traditional cinema. On the contrary, the audience (player) may be quite willing to undergo repeated experiences in order to remedy the regrets left from

their first experience. On the other hand, this aesthetic interaction is not merely simple "repetition" but is always situated within a "dynamic" process. Excellent interactive film and television works do not only offer choices at plot nodes, but intersperse choices throughout ordinary actions and dialogues. Moreover, it is often not a single choice that brings about a specific ending, but a multitude of intersecting choices accumulated over a long period that ultimately leads to a certain outcome. As an audience (player), one can never know exactly when one has made a critical choice; one simply chooses according to one's heart, ultimately arriving at a particular ending. This concealment of "choice nodes" shrouds the work in an aura of uncertainty and keeps it perpetually in a state of "dynamism." Whenever the audience (player) undertakes a repeated experience, they tend to try an alternative option in some choices about which they were "ambivalent," only to discover in the end that it was precisely these "ambivalent" and "seemingly inconsequential" choices that led to a tremendous change in the final outcome. This dynamic, uncertainty-laden mode of experience is one of the defining characteristics of interactive film and television.

4.3 "Subjective Empathy" Aesthetic Interaction

Interactive cinema differs from traditional film and television in that its experience is accompanied by the investment of a great deal of subjective thought. In traditional film and television, our judgments of the content and our conjectures about the plot do not affect the content of the film itself; in essence, the audience (player) remains outside the film, waiting for it to deliver one surprise after another. The characteristic of interactive cinema lies in the fact that the audience's (player's) real-time judgments and conjectures about the film are directly reflected in the subsequent plot, producing an effect of immediate feedback, and all outcomes are the sole responsibility of the audience (player) themselves. This means that during the experience, the audience (player) must constantly participate in the story, engaging in "deduction," "analysis," and "judgment," choosing with whom to interact, choosing which words to speak and which not to speak, and selecting the options they believe to be correct. The great freedom of aesthetic interaction is accompanied by immense moral and logical pressure; the audience (player) fears the serious consequences that a wrong choice might bring. In the work *Life Is Strange*, a small town may be led to destruction because of the audience's (player's) choices, and the lives of countless people may thereby come to an end. The "psychological pressure" faced by the protagonist in the film is concretized through choices and placed upon the audience (player), who must constantly interact with themselves, asking: Are all these choices worthwhile? Is this the right path? This "dynamic," "internally generated" aesthetic interaction is precisely the greatest allure of interactive cinema. Compared with traditional cinema, the subjective empathic aesthetic experience in "interactive film and television" originates from the audience's (player's) ability to change everything. In *Life Is Strange*, when the protagonist returns to the past and changes events that have already occurred, it brings about a tremendous upheaval in present life. Just when the audience (player) believes they have changed the life of their friend Chloe through their unremitting efforts, an even greater crisis follows. Because this consequence is the result of the choices the audience (player) made during the viewing process, they naturally feel that they themselves have caused the misfortunes befalling the characters in the work. The barrier between reality and virtuality is broken, and driven by an immense sense of guilt, the audience (player) will begin to constantly try to make amends for everything they have caused. This kind of subjective empathy based on the power to change is, compared to the empathic effect in traditional film and television, clearly more profound and more deeply moving.

5. Conclusion

As a vanguard in the development of contemporary literature and art, film and television art represents the spiritual needs of the broad masses of the people. Today, film and television art has undergone tremendous changes, whether in its forms of expression, its social status, or the connotations it carries. Throughout the history of film and television development, technology has always been a powerful driving force. "Interactive film and television," as an entirely new mode of film and television expression that has emerged alongside modern technology, likewise possesses enormous space for growth. For over a century, film and television art has long been unable to break free from the predicament of one-way communication. The emergence of "interactive film and television" has expanded new directions of communication on the basis of traditional one-dimensional indoctrination, offering new possibilities for the all-new development of film and television art. In recent years, interactive film and television have developed rapidly, and different directions of development have begun to emerge. At this juncture where this nascent art form is swiftly rising, it is particularly important to provide theoretical direction and pathway references for the development of interactive film and television. However, it is foreseeable that as its development becomes increasingly diversified and popularized, new problems will also arise. This requires scholars to pay close attention to theoretical research on interactive film and television, to promote its development, to provide practical guidance for its sustained advancement, and to contribute theoretical models to the entire film and television industry.

References

Chen Baijun & Huang Xinyuan. (2024). A Study on the Gamification Design of Interactive Cinema. *Contemporary Cinema*, (1).

Gao Yuan. (2022). *Aesthetic Research on Digital Film Art*. Changchun: Jilin University.

H. R. Jauss. (1987). *Reception Aesthetics and Reception Theory.* Shenyang: Liaoning People's Publishing House, p. 24.

Laura Mulvey. (1975). *Visual Pleasure and Narrative Cinema*. Cologne. KOENIG BOOKS.

Ni Taile, Chen Yingshuang, Wu Jinlong. (2021). From Audience to Player: On the Evolutionary Path of Interactive Video. *Media*, (7).

Shao Meng. (2018). A Study of Film Space Aesthetics. Fuzhou: Fujian Normal University.

Sigmund Freud. (2004). *Three Essays on the Theory of Sexuality and On the Unconscious*. Changchun: Changchun Publishing House.

Tong Xiaoxu. (2023). A Study on the Online Community Communication of Domestic Drama Fan Groups. Nanjing: Nanjing University of the Arts.

Wenqi Liu. (2023). Aesthetic Trends in Digital Cinema in the New Era. *Acad. J. Hum. Soc. Sci.*, (6), 61-65.

Zhang Dewei. (2020). A Preliminary Exploration of Innovative Paths for Interactive Video in the Context of 5G Technology. *West China Broadcasting TV*, (6).

Paradigm Academic Press
Art and Society
ISSN 2709-9830
JUN. 2026 VOL.5, NO.2

A Review of Aesthetic Assessment, Preference Alignment, and Explanation in Text-to-Image Generation

Beibei Nan[1]

[1] Zhengzhou University, Zhengzhou, Henan, China

Correspondence: Beibei Nan, Zhengzhou University, Zhengzhou, Henan, China.

doi:10.63593/AS.2709-9830.2026.06.005

Abstract

This study presents a bibliometric analysis complemented by a narrative review to systematically examine generative computational aesthetics in text-to-image (T2I) generation from the perspective of a closed-loop optimization framework. Specifically, it investigates the types and functions of aesthetic signals and their interfaces with generative model optimization. Aesthetic signals are categorized according to their output bandwidth into scalar scores, rating distributions, multidimensional attribute labels, natural-language critiques, and saliency maps/spatial attributions. Their respective roles in post hoc reranking, reward modeling and post-training alignment, multi-objective optimization, inference-time control and local editing, as well as personalization and interactive interfaces are further analyzed. Building upon this framework, the study examines the effectiveness of explanation mechanisms within the closed loop by evaluating existing approaches from four complementary dimensions: localizability, verifiability, executability, and causal reliability. The review indicates that although substantial progress has been achieved in individual components of aesthetic assessment, preference alignment, and explainability, a fully integrated aesthetics–alignment–explanation closed-loop paradigm remains largely underexplored. This review therefore provides a unified perspective for understanding and advancing closed-loop optimization in T2I generative systems.

Keywords: text-to-image generation, computational aesthetics, aesthetic assessment, preference alignment, explainability

1. Introduction

The rapid advancement of generative image models has shifted the focus of computational aesthetics from merely improving image quality to aligning generated outputs with human aesthetic preferences. Within this context, increasing attention has been devoted to a closed-loop paradigm that integrates aesthetic assessment, preference alignment, and explanation through learnable aesthetic signals. Rather than serving solely as a post hoc evaluation mechanism, aesthetic assessment is expected to continuously generate informative supervisory signals that guide generative models toward human preferences. Explanatory modules further diagnose generation failures and provide feedback for subsequent optimization, enabling the system to enter a new round of evaluation. Ideally, these three components operate collaboratively to establish a self-improving closed-loop framework in which generated images are iteratively refined.

Although various feedback mechanisms have been incorporated into recent T2I systems, most existing studies remain limited to isolated interfaces rather than forming a mature evaluation–alignment–explanation–re-evaluation paradigm. Depending on the depth at which feedback is incorporated into the generation pipeline, existing approaches can generally be categorized into three representative pathways.

The first pathway is the evaluation–selection interface, in which aesthetic assessment or preference models are

employed for candidate reranking, filtering, rejection sampling, or post-training. Representative approaches include ImageReward (Xu J et al., 2023), Human Preference Score (HPS) (Wu X et al., 2023), and Pick-a-Pic/PickScore (Kirstain Y et al., 2023). These methods efficiently integrate learned evaluators into reranking or post-training pipelines but generally compress aesthetic judgments into scalar preference signals, providing limited insight into why a generated image is preferred. Consequently, their interpretability and diagnostic capability remain relatively weak.

The second pathway is the evaluation–diagnosis interface. Representative techniques include Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju RR et al., 2017), Testing with Concept Activation Vectors (TCAV) (Kim B et al., 2018), together with more recent question-answering-based inspection frameworks and saliency-map-based diagnostic methods. These approaches primarily aim to improve the interpretability of generated images by revealing model attention or concept activation. Nevertheless, most of them function only as post hoc diagnostic tools and rarely provide actionable feedback that can be directly incorporated into model optimization.

The third pathway is the diagnosis–correction interface, where explanatory feedback is further transformed into optimization signals for local editing, image regeneration, sample filtering, training objectives, or iterative refinement during inference. For example, Rich Human Feedback (RichHF) (Liang Y et al., 2024) introduces region-level and token-level human annotations to facilitate sample filtering and masked image correction. Focus-N-Fix (Region-Aware Fine-Tuning for Text-to-Image Generation) (Xing X et al., 2025) converts detected local generation errors into region-aware fine-tuning objectives, while the recently proposed RationalRewards framework (Amodei D et al., 2016) first generates multidimensional critiques and subsequently employs reward learning together with iterative refinement during both training and inference. These studies demonstrate that explanation is evolving from a mechanism for post hoc interpretation into an effective optimization interface. Nevertheless, they generally concentrate on individual stages of the feedback loop and seldom provide a unified theoretical framework or systematic empirical validation.

Accordingly, this review focuses on text-to-image generation, one of the most representative applications of generative artificial intelligence, and reorganizes existing studies on assessment, alignment, and explanation through the lens of aesthetic signals. A closed-loop analytical framework is proposed to examine current methodologies and identify future research challenges. Following the PRISMA 2020 guidelines, a bibliometric analysis based on the Web of Science Core Collection (WoS) is conducted to investigate publication trends, thematic clusters, and the evolution of research topics. Based on this framework, the following research questions are proposed.

RQ1. What publication trends, thematic clusters, and topic evolution characterize research on computational aesthetics in text-to-image generation?

This question aims to provide a macroscopic overview of how research on computational aesthetics for T2I generation has evolved over time, identify its major research themes and knowledge structures, and establish an overall conceptual map for the subsequent review of aesthetic assessment, preference alignment, and explanation.

RQ2. What constitutes aesthetic signals in text-to-image generation systems?

This question investigates different forms of aesthetic signals — including scalar scores, rating distributions, attribute labels, and natural-language critiques — and analyzes their suitability for downstream tasks such as reranking, diagnosis, controllability, and feedback.

RQ3. How are aesthetic signals transformed into optimization objectives?

Once aesthetic preferences are incorporated into the generation process through reward models, parameter updates, or inference-time guidance, they inevitably introduce different benefits and trade-offs. This question therefore examines where aesthetic signals enter the optimization pipeline and how their integration influences system performance.

RQ4. What constitutes an effective explanation in computational aesthetics for text-to-image generation?

Saliency maps, concept-based explanations, and attention-attribution analyses of diffusion models provide evidence at different levels of abstraction. However, richer semantic explanations do not necessarily imply greater reliability. The central concern is whether explanations can be effectively incorporated into the closed loop to support controllable generation, preference alignment, and feedback-driven optimization.

2. Literature Sample Construction and Closed-Loop Topic Mapping

This chapter examines the empirical basis of the proposed closed-loop interface framework by identifying the distribution relationships of assessment, alignment, and explanation in the literature network based on the WoS core sample identification. It answers RQ1: What structural characteristics does the research on text-generated

graph computational aesthetics exhibit in terms of literature growth trends, topic clustering, and topic evolution? Can these characteristics support the subsequent chapters' development around a closed loop of aesthetic signal representation, optimization transformation, and interpretive feedback?

2.1 Data Sources and Screening Strategies

According to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA & Brennan SE., 2021; Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA & Brennan SE, 2021), all data used in the bibliometric analysis came from WoS, with a search date of May 7, 2026. WoS was chosen as the single data source because its bibliographic data is structured, its fields are consistent, and it provides citation links, facilitating annual trend, co-occurrence network, and topic analysis. If cross-database mixed searches are conducted, problems such as inconsistent citation formats and duplicate metadata may occur, increasing the workload of data cleaning (Donthu N et al., 2021; Zupic I & Čater T., 2015). Since this paper focuses on text-based image generation systems, the search terms revolve around three core issues: first, aesthetic evaluation in text-based image generation; second, preference learning, reward modeling, and preference alignment; and third, interpretability, attention attribution, and visual interpretation. The search query can be summarized as: TS = [(G NEAR/10 A) OR (G NEAR/10 P) OR (G NEAR/10 E)] NOT TS = X. For a detailed terminology set, please refer to Table 1. In addition to the above search query conditions, this paper did not further restrict the search by subject category, research direction, or source journal in WoS to avoid potential omissions caused by database classification rules.

Table 1. Search strategy for the WoS-based core corpus

Term set	**Meaning**	**Representative terms**
G	Text-to-image generation	text-to-image, text to image, image generation, diffusion model, latent diffusion
A	Aesthetic assessment	computational aesthetics, image aesthetics, aesthetic assessment, aesthetic evaluation, aesthetic quality
P	Preference alignment	human preference, preference learning, preference alignment, reward model, human feedback, RLHF, DPO
E	Explainability	explainability, interpretability, attention attribution, saliency map, concept explanation, visual explanation
X	Exclusion terms	IQA, super-resolution, denoising, style transfer, medical imaging, HCI, interior design, music, dance, philosophy

The initial search yielded 172 records. Subsequently, the search results were manually screened (inclusion and exclusion criteria are detailed in Table 2), ultimately retaining 53 core documents for bibliometric statistics and keyword analysis. The flowchart is shown in Figure 1. Considering that some high-impact achievements in the computer science field may not be fully covered by WoS, this paper further supplemented the descriptive review section with searches of datasets such as arXiv, Scopus, and IEEE Xplore to identify potentially overlooked key methods and latest advancements. The supplementary search results were primarily used for the qualitative discussions in Chapters 3 to 5 and were not included in the bibliometric statistics of the WoS core sample to maintain the consistency and reproducibility of the data sources for the bibliometric analysis.

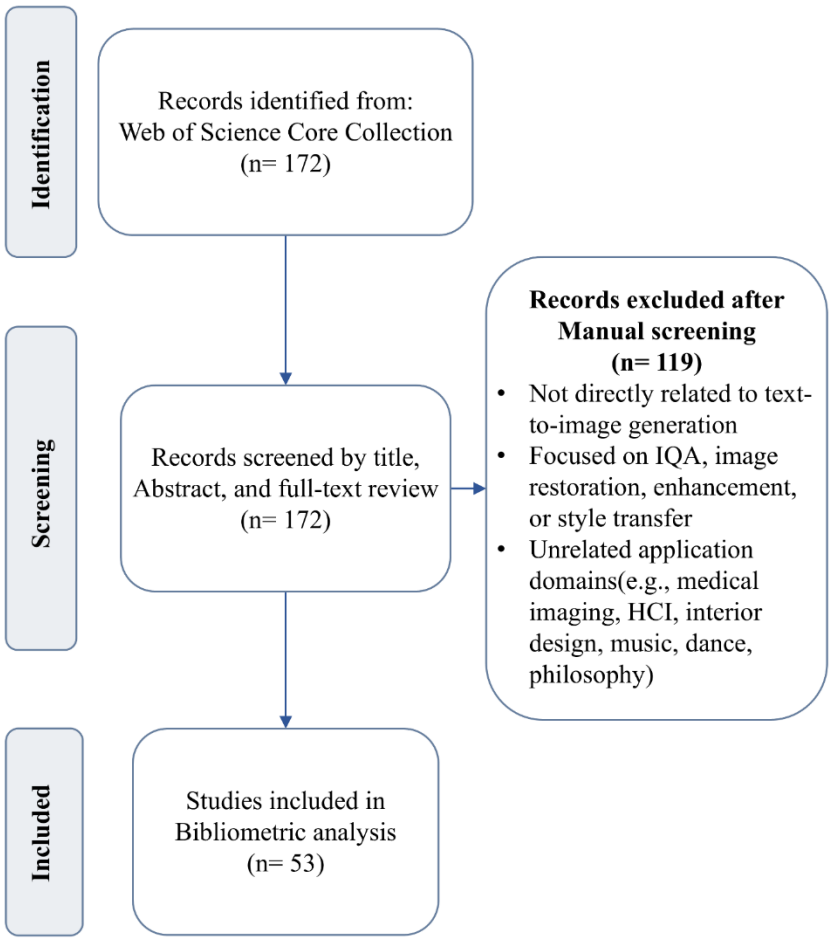

Figure 1. PRISMA 2020-based flow diagram of study selection

Table 2. Inclusion and exclusion criteria

Criterion	Inclusion	Exclusion
Research object	T2I generation, text-guided image synthesis, diffusion-based T2I	Non-T2I generation, general image editing, image-to-image translation
Research focus	Aesthetic assessment, preference alignment, reward modeling, explainability	IQA, restoration, enhancement, super-resolution, denoising
Domain	Computational aesthetics and generative AI	Medical imaging, HCI, interior design, music, dance, philosophy
Document type	Article, Proceedings Paper, Review Article	Editorial, news, book review, irrelevant conference abstract
Language	English	Non-English

2.2 Data Cleaning

Before conducting the bibliometric analysis, this paper first performed necessary data cleaning on the 53 data points exported from WoS, standardizing synonyms and spelling variations. For example, text-to-image, text to image, and text-to-image synthesis were unified as text-to-image generation; diffusion models, latent diffusion, and stable diffusion were unified as diffusion model; and direct preference optimization (DPO) and DPO were unified as DPO. Simultaneously, generalized terms lacking clear thematic focus, such as model, method, framework, system, image, and generation, were removed to improve the interpretability of the keyword map.

2.3 Results and Discussion

Bibliometric analysis of computational aesthetics in text-to-image generation allows us to analyze the number of

publications, visually demonstrating the significant growth in scientific output, as shown in Figure 2. The analysis reveals a marked increase in activity in this field after 2023, culminating in explosive growth in 2025. Furthermore, although 2026 is not a complete year, a continued upward trend in quantity is already evident. This trend not only reflects quantitative growth but also suggests that the field is shifting from simple model capability research to issues such as aesthetic evaluation, preference feedback, and alignment optimization.

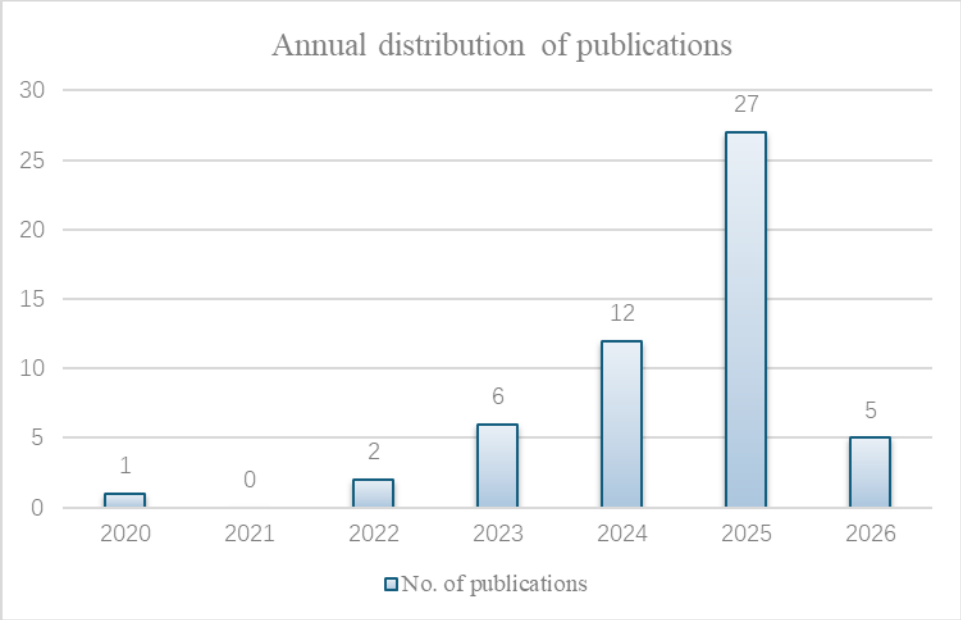

Figure 2. Annual distribution of publications

Figure 3 illustrates the keyword co-occurrence network of the core samples, identifying six distinct topic clusters. The yellow T2I generation and red diffusion model are the most prominent central nodes, illustrating that text-to-image generation and diffusion models constitute the fundamental technological background of this field. Around these two central nodes, the network further differentiates into several interconnected topic directions, though the connections between them remain loose. One is the aesthetic assessment direction, represented by aesthetic assessment, human preference, and vision-language alignment; the second is the preference alignment direction, represented by preference alignment, preference learning, human feedback, reward model, DPO, and reinforcement learning; and the third is the explanatory analysis direction, represented by explainability and class activation mapping.

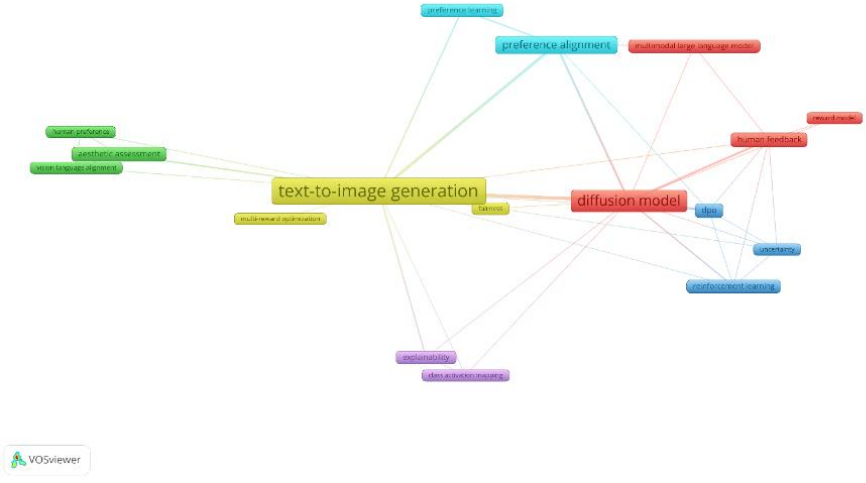

Figure 3. Keyword co-occurrence network of the T2I computational aesthetics corpus

Building upon the keyword contribution network, this paper further presents a temporal overlay distribution map of keywords, as shown in Figure 4. Unlike the previous keyword co-occurrence map, the temporal overlay map uses color variations to display the average year of occurrence of different keywords, thereby helping to identify recently active topics. In the figure, newer keywords are mainly concentrated in areas such as DPO, reinforcement learning, uncertainty, and multi-reward optimization, indicating that interpretive analysis and multimodal language feedback are beginning to enter the research framework of text-based image aesthetics.

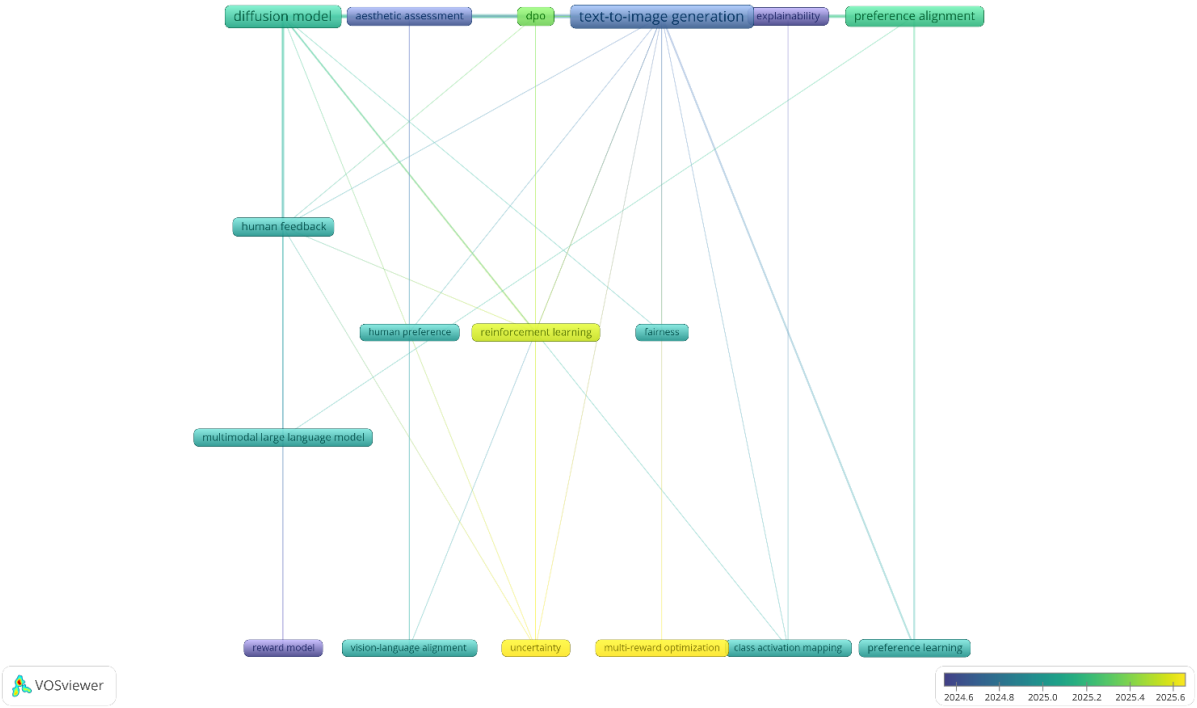

Figure 4. Temporal overlay map of keywords in the T2I computational aesthetics corpus

Finally, based on the keyword co-occurrence and temporal superposition analysis, and combined with the research questions of this paper, the core keywords are further organized into three functional modules: assessment, alignment, and explanation, as shown in Figure 5. The central part is T2I Computational Aesthetics, representing the task scenario this paper focuses on. In the assessment module, scores and attribute labels in aesthetic assessment are suitable for ranking and diagnosis; human preference can reflect subjective judgments at the user level; vision-language alignment expands the relationship between aesthetic evaluation and semantic consistency; and uncertainty reminds researchers that aesthetic judgment itself has distribution and uncertainty. In the alignment module, keywords such as DPO, reinforcement learning, and multi-reward optimization indicate that aesthetic signals are no longer just posterior evaluations, but are gradually being transformed into optimizable objective functions, reward models, or training strategies. In the explanation module, explainability, class activation mapping, and multimodal large language model point to the trend of interpretability of aesthetic feedback. Therefore, this three-module framework transforms the bibliometric findings into the analytical structure of the subsequent review, placing the three in a closed-loop system.

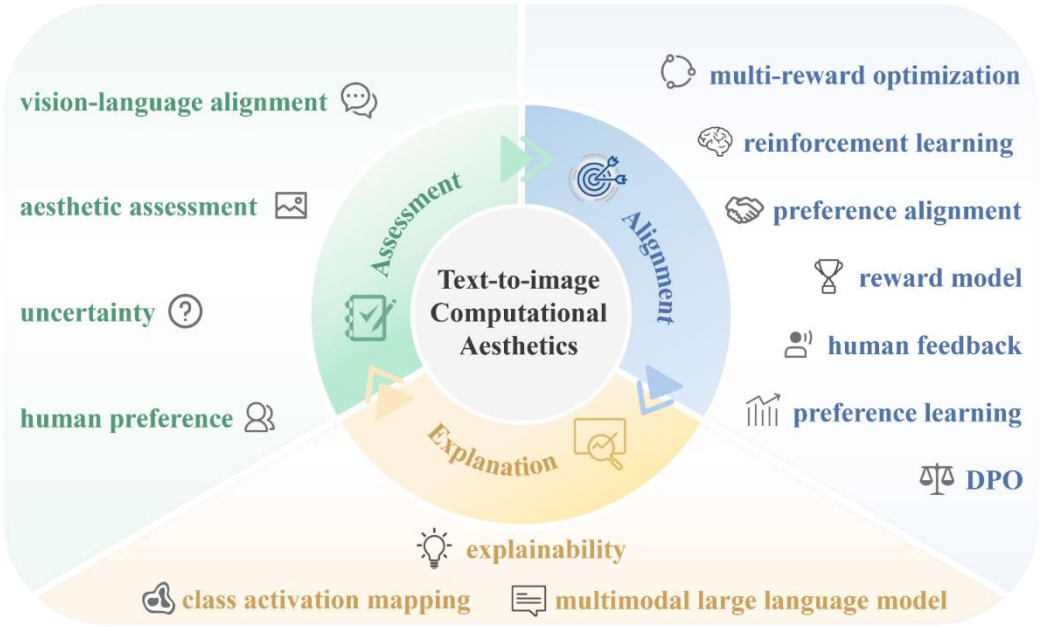

Figure 5. Keyword-informed conceptual framework of T2I computational aesthetics

2.4 Summary

Based on the bibliometric analysis of the above 53 core samples, we can see that text-to-image generation computational aesthetics is a relatively new, rapidly growing interdisciplinary research direction with a limited sample size. Annual publication trends indicate that this field accelerated significantly after 2023, reaching its peak in 2025. Keyword co-occurrence networks show that T2I generation and diffusion models constitute the technological center, while aesthetic evaluation, preference alignment, and explanatory analysis form three main functional branches. The time-overlay plot further indicates that recent research focus is shifting towards human feedback, DPO, reward models, multi-reward optimization, and explainability. However, the connections between keywords also show that most existing research remains at the level of local interfaces or pairwise couplings, failing to form a stable "evaluation – alignment – interpretation – re-evaluation" closed loop. Therefore, subsequent chapters of this paper will discuss the closed-loop interfaces in sequence: how aesthetic signals are represented, how they are transformed into optimization objectives, and how interpretation becomes verifiable and executable feedback.

However, this paper also acknowledges certain limitations of the bibliometric sample. First, WOS, as a single database, cannot fully cover all articles in the field of computer science. Second, the selection of keywords may have affected the completeness of the research, excluding emerging or less common terms that could enrich the analysis. Third, manually filtering from 172 original documents to 53 may have resulted in the omission of some documents. Fourth, the final core sample is relatively small, and some records lack author keywords and Keywords Plus; therefore, the keyword map should be understood as a structural overview of the focused sample, rather than a macro-level knowledge graph of the entire generative AI field. Based on this, this paper further supplements the key methods and latest advancements in the subsequent descriptive review with sources such as arXiv, Scopus, and IEEE Xplore to enhance the completeness and cutting-edge nature of the qualitative analysis.

3. Aesthetic Signals in Text-Based Image Generation Systems

In text-based graph generation systems, aesthetic signals cannot be simply understood as a single aesthetic score. More accurately, they are the computable expression of human aesthetic judgment of the generated results within the system. Depending on the output form, they can be represented as scores, preference pairs, and reward values, or as rating distributions, uncertainty weights, attribute labels, natural language criticism, or spatial attribution graphs. The key differences between these signals lie in their information bandwidth, degree of structuring, and accessible system interfaces. Therefore, this chapter addresses the questions raised in RQ2 by exploring the output forms of aesthetic signals in text-based graph systems and their suitable downstream task interfaces (see Table 3).

In traditional image aesthetic evaluation, aesthetic signals usually appear in the form of average scores or rating distributions, such as AVA (Murray N et al., 2012) and NIMA (Talebi H & Milanfar P., 2018), which represent a

shift from average scores to rating distribution modeling. However, based on the WoS core sample and supplementary retrieval, it can be seen that in the text-based image generation system, the function of aesthetic signals has been extended from posterior evaluation to generation loop: it is no longer just a posterior evaluation of the generation result, but may further enter the generation system loop for reordering, sampling selection, reward modeling, preference alignment, local correction and interactive feedback. For example, ImageReward Human preferences are transformed into reward models; (Xu J et al., 2023) VIEScore (Visual Instruction-guided Explainable Score) uses a multimodal large language model in conditional image generation evaluation (Ku M et al., 2024), so that the evaluation results have natural language interpretation capabilities.

Table 3 summarizes the main types of aesthetic signals and their interface functions in the text-to-image closed loop. It should be noted that the classification objects in the table are not single documents, but rather the output forms of aesthetic signals. A single study may simultaneously contain multiple signals such as ratings, attributes, textual feedback, or spatial attribution; therefore, this paper summarizes them according to their main functions in the closed loop.

Table 3. Closed-loop interface function of aesthetic signals in the Wenshengtu system

Signal level	Typical form	Closed-loop interface function	Main limitations	Representative studies
Low bandwidth optimized signal	Score, preference pair, reward, win rate, binary feedback	Sorting, reordering, sampling selection, reward modeling	Compressed evaluations are prone to proxy bias.	ImageReward (Xu J et al., 2023); HPS (Wu X et al., 2023); Pick-a-Pic/PickScore (Kirstain Y et al., 2023); Diffusion-KTO (Ethayarajh K et al., 2024)
Robustness and Coordination Signals	Rating distribution, uncertainty weights, multi-reward / multi-preference structure	Reward calibration, risk control, sample weighting, and multi-objective coordination	It is difficult to provide the cause of local corrections, and the calibration is complex.	NIMA (Talebi H & Milanfar P., 2018); Calibrated Multi-Preference Optimization (Lee K et al., 2025); CAPO (Hu S., 2026); Flow-Multi (Lee J & Choi J., 2026); Parrot (Lee SH et al., 2024)
Diagnostic signals	Diagnostic dimensions include attribute labels, multidimensional evaluation, semantic / quality / realism, etc.	Target decomposition, error diagnosis, and condition control	Reliance on tagging systems leads to high annotation costs.	Learning Multi-Dimensional Human Preference (Zhang S et al., 2024); MindScore (Tong Y et al., 2025); AGIAA-2K (Hu B et al., 2025)
Feedback signals	Natural Language Critique, QA Interpretation, Saliency Map, Region Mask	Feedback correction, partial editing, interactive evaluation	It needs to be structured; there may be hallucinations or attribution biases.	VIEScore (Ku M et al., 2024); TIFA (Hu Y et al., 2023); Visual Programming (Cho J et al., 2023); DAAM (Tang R et al., 2023); RichHF (Liang Y et al., 2024); TNIM (Han S-H & Choi H-J., 2025); Sal-Guide Diffusion (Lin X et al., 2024)

3.1 Low-Bandwidth Optimized Signal

Low-bandwidth signals such as scalar ratings and preference rewards are the most common aesthetic signal types in current text-based graph systems. Their typical forms include overall aesthetic score, preference score,

win rate, reward value, or one-dimensional utility value. For generative systems, these signals are the easiest to enter into sorting and can directly enter the sampling selection, reward fine-tuning, DPO, or reinforcement learning process, making them the most effective optimization interface. ImageReward (Xu J et al., 2023), HPS (Wu X et al., 2023), and Pick-a-Pic / PickScore (Kirstain Y et al., 2023) show that human preferences can be compressed into callable ratings; Diffusion-KTO (Aligning Diffusion Models by Optimizing Human Utility) (Ethayarajh K et al., 2024) further shows that even if the feedback is just a minimalist second-order feedback signal likes / dislikes, it can become an effective alignment signal. Another noteworthy variation of scalar rating research is the FERGI (Automatic Annotation of User Preferences for Text-to-Image Generation from Spontaneous Facial Expression Reaction) model proposed by Feng et al. (2025). It automatically generates user preference ratings by collecting spontaneous expressions when users view generated images, demonstrating that the collection of scalar signals does not necessarily depend on explicit questionnaires. This idea has certain practical implications for human-computer interaction systems, but it also involves new issues such as privacy, scene noise, and cross-user differences.

These studies collectively drive the shift in aesthetic evaluation of text-based images towards reward signals driven by human preferences. However, because scalar ratings and preference rewards compress factors such as aesthetics, visual quality, semantic consistency, composition, and detail into a single numerical value, it becomes difficult to ascertain the specific reasons for high or low scores. Furthermore, using a single numerical rating as a training objective can easily lead to reward hacking or reward overfitting, where increasing the score does not necessarily improve aesthetic quality. Therefore, these signals are suitable as optimization entry points within a closed loop, rather than as a complete aesthetic explanation.

3.2 Robustness and Coordination Signals

Compared to scalar scores and preference rewards, rating distributions, uncertainty signals, and multi-preference structures attempt to preserve the discrepancies in human aesthetic judgments. By outputting probability histograms of human ratings, they explicitly characterize the subjective uncertainty and degree of controversy in human aesthetic judgments. NIMA As a foundational work of related methodologies (Talebi H & Milanfar P., 2018), it predicts the distribution of human ratings in the form of histograms. Just as in text-based graph systems, when the model generation faces different users, different cultural backgrounds and different usage scenarios, a single average score is difficult to fully express the uncertainty of aesthetic judgment. Especially in open text-based graph tasks, the same prompt may correspond to multiple reasonable visual implementations, and aesthetic preferences themselves are distributed rather than unique.

In recent studies of Wensheng graphs, this idea has been transformed more into uncertainty modeling, reward calibration, multi-preference structure and multi-objective reward optimization. Calibrated Multi-Preference Optimization focuses on the calibration problem between multiple reward models and handles the conflict between multiple preference objectives through expected win-rate and Pareto-frontier pair selection (Lee K et al., 2025). In contrast, Causal-Adaptive Preference Optimization (CAPO) (Hu S., 2026). This paper distinguishes between aleatoric uncertainty and epistemic uncertainty from the perspective of causal uncertainty, and dynamically adjusts the preference optimization loss accordingly to reduce the impact of noisy preference pairs and enhance learning on difficult samples. Flow-Multi (multiflow framework combining multiple generative flows for enhanced alignment) (Lee J & Choi J., 2026) and ParrotParrot (Personalized Response Optimization Through Reinforcement and Offline Training) (Lee SH et al., 2024) further organize multiple evaluation objectives into multidimensional reward or multi-objective optimization problems, and coordinate aesthetics, human preference, text-image alignment, and GenEval (An Object-Focused Framework for Evaluating TexttoImage Alignment) through Pareto dominance or Pareto-optimal selection, (Ghosh D, Hajishirzi H & Schmidt L., 2023) or trade-offs between objectives such as image sentiment. These works do not directly generate rating distributions, but collectively show that textual image evaluation and alignment are shifting from a single scalar score to a signal structure with multiple preferences, multiple rewards, and uncertainties. From the downstream interface, this signal type is suitable for connecting evaluation and alignment, but it still cannot identify the specific direction of correction in local image areas.

3.3 Diagnostic Signals

With single-rating signals, the advantage of multi-dimensional attribute signals lies in their diagnostic nature. It can break down the overall judgment of “good or bad” into diagnosable sub-problems, such as style, emotion, composition, missing subject, local mismatch, artifacts, detail quality, etc., providing clearer targets for subsequent control and correction. In text-to-image systems, Learning Multi-Dimensional Human Preference for Text-to-Image Generation is a representative work in this direction (Zhang S et al., 2024). This study breaks down human preferences into dimensions such as aesthetics, semantic alignment, detail quality, and overall assessment, so that preference judgment is no longer just a binary choice, but has an interpretable multi-dimensional structure. (Tong Y et al., 2025) MindScore also quantifies text-to-image preferences from the

perspectives of matching, faithfulness, quality, and realness. In addition, (Hu B et al., 2025) AGIAA-2K (A Fine-grained Dataset for Aesthetic and Alignment Evaluation of AI-Generated Images) constructs a fine-grained evaluation resource for Artificial General Intelligence (AGI) aesthetics and alignment at the dataset level, indicating that attribute-based supervision is shifting from single-preference comparison to a more fine-grained evaluation space.

However, attribute labels and multidimensional aesthetic signals also have limitations that cannot be ignored. First, the label system itself will presuppose which dimensions have evaluation value. If the label space is designed too narrowly, the model may only optimize around the predetermined dimensions and ignore the implicit factors that are difficult to name but affect aesthetic judgment. Second, fixed attribute labels are difficult to fully cover the differences in aesthetic experience when facing user groups with different cultural backgrounds and aesthetic styles. Third, excessively fine label granularity will also bring sparsity and cost problems. (Collins KM et al., 2024) As Collins et al. have shown, although fine-grained feedback is more advantageous in meeting diverse social preferences, its cost and the complexity of building the model limit its implementation. Therefore, attribute labels are more suitable as diagnostic and control interfaces rather than the final representation of aesthetic judgment.

3.4 Feedback Signals

Natural language criticism and spatial attribution maps provide aesthetic signals with higher bandwidth than attribute labels. They not only indicate whether an image is good or bad, but also attempt to pinpoint the specific semantic, object, region, or generation stage where the problem occurs. Therefore, they are suitable for inpainting, local resampling, interactive editing, data cleaning, and human-computer collaborative evaluation. Compared to scalar scores, they offer greater operability; compared to attribute labels, they preserve more contextual and spatial information.

Representative works in natural language criticism mainly include VIEScore (Ku M et al., 2024), Text-to-image Faithfulness Evaluation with Question Answering (TIFA) and Visual Programming (Hu Y et al., 2023; Cho J et al., 2023). Among them, (Ku M et al., 2024) VIEScore uses a multimodal large language model to perform interpretable evaluation of conditional image generation tasks, attempting to make the evaluation result not just a score, but include natural language judgments and reasons. (Hu Y et al., 2023) TIFA focuses on analyzing the faithfulness of generated images at the semantic and detail levels. (Cho J et al., 2023) Visual Programming for Text-to-Image Generation and Evaluation introduces the interpretable stepwise T2I generation framework Visual Programming Generation Framework (VPGen) and proposes the visual programming-based T2I generation interpretable evaluation framework Visual Programming Evaluation Framework (VPEval), which makes object generation, layout control and evaluation results have clearer intermediate states. These studies together show that the value of natural language signals lies in their ability to preserve the relationship between failure types, target words, reasoning processes and modification cues.

Spatial attribution and saliency signals further locate the feedback to the image region or text token. Diffusion Attentive Attribution Maps (DAAM) is an important starting point in this direction (Tang R et al., 2023). It generates a word-to-pixel attribute map by aggregating cross-attention of Stable Diffusion and studies lexical and semantic phenomena accordingly. (Liang Y et al., 2024) The study of Liang et al. requires annotators to point out the areas in the image that do not match or are misaligned with the text, and to point out which words in the text are incorrectly represented or missing in the image, thereby achieving more granular human feedback supervision. (Han S-H & Choi H-J., 2025) TNIM combines cross-attention mechanism and interpretability method to generate heatmaps and reveal how a single text tag affects the image generation process in multiple time steps. (Lin X et al., 2024) The Sal-Guide Diffusion (Saliency Maps Guide Emotional Image Generation through Adapter) method focuses saliency maps on the generation of emotional images. By introducing saliency information into the diffusion process through the adapter, it significantly improves the generation effect of emotional images and also shows that spatial signals can not only be used for post-hoc interpretation, but also enter the generation guidance.

However, these high-bandwidth signals require further structuring to be stably incorporated into the training process. If these signals originate from a surrogate model rather than genuine human feedback, they may also inherit the surrogate model's illusions, biases, or attentional deviations. Therefore, high-bandwidth signals cannot replace scores or rewards, but rather supplement diagnostic, constraint, and corrective information that scores cannot provide.

3.5 Summary

In summary, the aesthetic signals in the system can be viewed as a continuous spectrum with bandwidth ranging from low to high. Low-bandwidth signals are easier to optimize but are prone to losing their underlying causes; medium-bandwidth signals can express divergences or decompose objectives but rely on calibration and labeling

systems; high-bandwidth signals are more diagnostic but difficult to directly convert into a stable loss function, and their structuring and verification costs are also higher. In other words, low-bandwidth signals are generally more suitable for entering the optimization interface, while high-bandwidth signals are more suitable for diagnostic, constraint, and feedback functions. They are not substitutes but rather different components forming the aesthetic closed loop. This means that aesthetic signals in T2I are better understood as a set of composable interfaces rather than a single indicator.

4. Transformation of Aesthetic Signals into Optimization Objectives in the Wensheng Image System

Building upon the answer in Chapter 3 regarding the aesthetic signals in the Wensheng graph system, this chapter further addresses RQ3: How are aesthetic signals transformed into optimization objectives? It aims to analyze the levels at which different signals enter the system, the ways they are transformed into objective functions, and the resulting benefits and costs.

In the T2I system, an aesthetic score can be used as a posterior evaluation index, or it can be trained as a reward model and further used for re-ranking, rejection sampling or model fine-tuning; a preference pair can be used for candidate result comparison, or it can be used for DPO, (Christiano PF et al., 2017) Reinforcement Learning from Human Feedback (RLHF) or other post-training objectives; a natural language critique can remain at the interpretation level, or it can be structured into question-answering constraints, preference samples, attribute labels or prompt word correction instructions; a saliency map can be used for visual diagnosis, or it can be transformed into region mask, adapter conditions or local inpainting control signals. Therefore, the transformation of aesthetic signals into optimization objectives is essentially a system interface design problem: the same signal will produce different optimization effects and risks in different interfaces.

From a system-level perspective, aesthetic signals enter the text-based graph generation process primarily through five types of interfaces: posterior selection interface, reward model and post-training interface, multi-objective coordination interface, inference-time control interface, and personalization and interactive alignment interface. These five interfaces correspond to transformation paths progressing from simple to complex: posterior selection only changes the output selection; reward model and post-training begin to influence parameter updates; multi-objective coordination handles reward conflicts; inference-time control focuses on local correction; and personalization and interactive alignment incorporates user preferences into the dynamic feedback process. These are not simply technological substitutions, but rather represent different trade-offs between optimization intensity, computational cost, interpretability, and risk control. Based on this framework, the following sections will analyze how aesthetic signals are transformed into optimization objectives through these interfaces.

Table 4 summarizes the main interfaces for transforming aesthetic signals into optimization targets in the image-based system. However, these interfaces are not mutually exclusive and the same method may span multiple interfaces. For example, (Xu J et al., 2023) ImageReward can be used for posterior reordering and can also be used for post-training through Reward Feedback Learning (ReFL); (Liang Y et al., 2024) RichHF can provide fine-grained feedback data and also support local repair.

Table 4. The main interface from aesthetic signals to optimization objectives in the Wensheng image system

Interface layer	Signal conversion method	Function	Main limitations	Representative literature
Posterior reordering	Sort and filter candidate images using scores, preference scores, or rewards	Low cost and flexible deployment make it suitable for quickly filtering results	You can only choose from existing results; it's difficult to change the generation mechanism	ImageReward (Xu J et al., 2023); HPS (Wu X et al., 2023); Pick-a-Pic/PickScore (Kirstain Y et al., 2023)
Reward Model and Post-Training	Preference pairs, ratings, or binary feedback can be incorporated into training objectives such as ReFL, DPO, and RLHF/RL	Directly affects parameter updates, resulting in more durable alignment effects	Susceptible to reward hacking, overfitting, and proxy bias	ImageReward (Xu J et al., 2023) /ReFL; Diffusion-DPO (Wallace B et al., 2024); Diffusion-KTO (Ethayarajh K et al., 2024)

Multi-objective coordination	Multiple rewards can be coordinated through reward calibration, uncertainty weighting, or Pareto selection	Improving robustness under multiple preference conflicts	The process is complex, and there is a lack of unified standards for balancing objectives	Calibrated Multi-Preference Optimization (Lee K et al., 2025); CAPO (Hu S., 2026); Flow-Multi (Lee J & Choi J., 2026); Parrot (Lee SH et al., 2024)
Inference-time control and local editing	Transform language criticism, QA, saliency maps, or region masks into prompts, adapters, or inpainting conditions	Flexible, suitable for local corrections and interactive feedback	Relying on the reliability of feedback makes it difficult to change basic preferences.	TIFA (Hu Y et al., 2023); VIEScore (Ku M et al., 2024); RichHF (Ethayarajh K et al., 2024); Sal-Guide Diffusion (Lin X et al., 2024)
Personalized and interactive alignment	Use user preference examples, multi-turn feedback, or preference embeddings as conditions	Support for personalized aesthetic generation	High data demand, with privacy and cold start issues	Personalized Preference Fine-tuning; Preference Adaptive and Sequential T2I; Imaginique Expressions

4.1 Posterior Reordering: Weak Alignment Interface of the Output Layer

Posterior re-ranking is the shallowest interface for aesthetic signals to enter the text-based image system. It scores, ranks, and filters candidate results using aesthetic rating systems, preference models, or regularized evaluation metrics. Unlike post-training alignment, posterior re-ranking typically does not directly modify the generative model parameters but rather applies external selection pressure after the results are generated.

At this level, scalar scores and aesthetic signals of preference rewards are most suitable. (Xu J et al., 2023) The importance of ImageReward is first reflected in that it trains human preferences into callable text-based image reward models, enabling the generated results to be automatically evaluated and candidate-selected according to human preferences. (Wu X et al., 2023) HPS, Pick-a-Pic and PickScore also reflect similar paths (Kirstain Y et al., 2023): they transform real user selections or pairwise preferences into scalable scoring functions, thereby supporting model comparison, candidate image re-ranking, and low-cost automatic evaluation. However, posterior re-ranking also has limitations. First, it can only select the relatively optimal results from the candidate set and cannot fundamentally change the model generation distribution. If there are no works that match the preferences in the candidate images, the scorer cannot create better results out of thin air. Second, this method requires generating multiple samples (best-of-N), which has a high sampling cost. More importantly, if the scorer itself is biased, the re-ranking will amplify this bias: the system will repeatedly select styles that match the scorer's preferences, ignoring the possibility of diversity. Therefore, posterior reordering is essentially a weak alignment interface that can improve the surface performance of the output distribution, but has limited ability to correct the generation mechanism. Its role is closer to output layer filtering than aesthetic learning at the generation mechanism level.

4.2 Post-Training Alignment: From Reward Model to Parameter Update

Compared to the post-test stage, the reward model and post-training alignment interface bring aesthetic signals into the parameter optimization update of the model. At this level, researchers usually train a callable reward model using scalar scores or preference pairs, and then integrate the model into the actual training process. The common paths include three categories: first, using the scorer as an external reward and optimizing the generative model; second, converting preference pairs into DPO or pairwise loss; and third, connecting the reward model to the post-training process of reinforcement learning or approximate reinforcement learning. For example, the key significance of ImageReward is that it not only learns the preference model (Xu J et al., 2023), but also connects the reward model to the ReFL fine-tuning of the diffusion model through Reward Feedback Learning, making the preference score the actual training target. Similarly, (Wallace B et al., 2024) the Diffusion-DPO method directly uses human preference pairs for DPO loss in post-training. If reinforcement learning (such as policy gradient with KL constraints) is used, the signal can be used as an external reward for optimization, like DPOK (Direct Preference Optimization with Knowledge-aware refinement) or DDPO (Wallace B et al., 2024).

At this level, scalar scores can be used for regression loss, preference pairs for pairwise loss or DPO objectives,

and reward models can seamlessly replace supervised objectives to optimize the generator. Therefore, generative models can truly learn aesthetically pleasing parameter settings, resulting in more persistent alignment. However, the trade-off lies in a more complex and problematic optimization process. First, a single compression of a low-bandwidth signal can lead to reward hacking: the model might exploit shortcuts beyond human intuition to improve scores. Second, direct fine-tuning or RL can cause overfitting, where the model overfits surrogate rewards, reducing diversity. Furthermore, noise and bias in the training objective are difficult to control. In summary, this path allows for deeper alignment with aesthetic goals, but requires careful balancing of the risks between optimizability and stability.

4.3 Multi-Objective and Uncertainty Optimization: From Single Reward to Objective Coordination

When aesthetic evaluation involves multiple objectives or there are scoring discrepancies, multi-objective coordination interfaces have emerged. These methods no longer force multiple objectives into a single score, but rather coordinate them at the optimization level. For example, (Lee K et al., 2025) Calibrated Multi-Preference Optimization calibrates the output distributions of multiple reward models and utilizes the Expected Win Rate (EWR). The strategy selects sample pairs to balance different sources of preference; (Hu S., 2026) CAPO distinguishes the types of uncertainty in the samples based on the causal model and dynamically adjusts the weights of each preference in the loss to avoid excessive interference of noise on training; (Lee J & Choi J., 2026; Lee SH et al., 2024) Flow-Multi and Parrot regard multidimensional preferences as vectors and use the Pareto dominance method to select samples within or between batches to take into account multiple objectives such as semantic accuracy, style preference, and technical quality.

The core of these methods lies in the reallocation of training samples and target weights, mitigating the contradictions of multi-dimensional signals through architectural design. Their main benefit is improved adaptability to diverse preferences; the model no longer needs to greedily optimize along a single reward dimension, but instead considers the balance of different metrics simultaneously. However, this also makes the training process more complex, requiring additional calibration steps and verification of the quality of the frontier solutions; the trade-offs between different aesthetic goals are also difficult to interpret; furthermore, if the reward model itself has significant bias, multi-objective schemes may simultaneously optimize multiple misleading signals.

4.4 Inference-Time Control and Local Editing: Structured Use of High-Bandwidth Signals

The key to the control interface during inference is not retraining the model, but structuring high-bandwidth feedback into executable constraints, such as prompt refinement, QA checks, region masks, adapter conditioning, or local inpainting. (Ku M et al., 2024; Hu Y et al., 2023) Methods such as TIFA and VIEScore mainly provide question-and-answer or linguistic diagnostics, which can be used to identify text - image inconsistencies and provide a basis for subsequent prompt word correction, sample filtering, or manual editing. (Cho J et al., 2023) Visual Programming further decomposes generation and evaluation into modular programs, allowing constraints such as objects, quantity, and layout to be explicitly checked. In contrast, (Liang Y et al., 2024; Lin X et al., 2024) RichHF and Sal-Guide Diffusion are closer to executable interfaces: the former transforms region-level and word-level feedback into masks and local repair cues, while the latter introduces saliency information into the diffusion process through the adapter. (Tang R et al., 2023; Han S-H & Choi H-J., 2025) DAAM and TNIM mainly provide attribution information for token-region or text-to-noise, providing an explanation basis for error location and subsequent control.

Therefore, the advantage of inference-time control lies in its flexibility and local operability, but its effectiveness is highly dependent on the structured quality of the feedback signal. If natural language criticism is unverifiable, saliency maps lack causality, or region mask localization is inaccurate, local control may actually introduce new errors.

4.5 Personalization and Interactive Alignment: From Group Preferences to User Preferences

Finally, the personalized and interactive alignment interface shifts aesthetic alignment from group preferences to user preferences, focusing on the subject differences in aesthetic preferences. Such methods believe that the same text prompt may correspond to multiple reasonable images, and different users have different aesthetic standards. Therefore, the typical approach of such methods is to inject user preferences as conditional signals into the model, rather than compressing all users into the same global reward model. (Dang M et al., 2025) Personalized Preference Fine-tuning of Diffusion Models learns personal preference embeddings through a small number of user preference samples and injects them into the cross-attention process of the diffusion model, and then combines them with the DPO goal to achieve personalized preference alignment. (Wan Y et al., 2024) Imaginique Expressions incorporates user personality and aesthetic preferences into the short text generation process through a personality encoder and a personalized aesthetic evaluation model. (Nabati O et al., 2024) Preference Adaptive and Sequential Text-to-Image Generation further extends preference alignment into a

multi-round interactive process, gradually approaching user intent through continuous prompt expansion and user feedback.

The advantage of this interface lies in its ability to generate customized results, better matching the aesthetic needs of the target users. However, it has high data requirements, needing to collect individual preference samples or feedback, and presents privacy and cold-start issues. More importantly, personalized alignment may weaken the model's public aesthetic benchmark, creating a new tension between "satisfying individual preferences" and "maintaining generalizable quality".

4.6 Summary

In summary, the transformation of aesthetic signals into optimization objectives is not a single path, but rather a spectrum of interfaces progressing from shallow to deep. Posterior reordering is at the shallowest level, only changing the output selection without altering the generation mechanism; reward models and post-training interfaces write preferences into parameter updates, enabling more persistent alignment, but are more susceptible to surrogate bias and reward hacking; multi-objective and uncertain optimization attempts to reconcile multiple inconsistent aesthetic objectives, improving robustness, but also increasing the difficulty of calibration and verification; inference-time control and local editing can utilize high-bandwidth signals for directional correction, but their effectiveness depends on the reliability of interpretation and feedback; personalized and interactive alignment incorporates user preferences into a closed loop, but also introduces privacy, cold-start, and generalization issues. Therefore, no single optimization method is optimal; rather, different aesthetic signals should enter different system interfaces based on their bandwidth, reliability, and objective granularity.

5. The Effectiveness of Interpretation in the Wensheng Graph Generation System

Chapter 4, which focuses on "how signals are transformed into objective functions," this chapter focuses on whether interpretive signals are sufficient to serve as evidence for feedback correction. In other words, before an interpretation can enter the closed loop, it must first be proven to be able to locate, verify, and execute. This chapter further answers RQ4: In the computational aesthetics of text-based graph generation, what kind of interpretation is effective? The key to this question is not to rank the various forms of interpretation, but to examine whether the interpretation can return to the closed loop and provide an operational basis for generation control, preference alignment, and feedback correction.

In traditional interpretability research, explanations are often understood as visual or verbal descriptions of the model's decision-making process, such as saliency heatmaps, concept activation, attention attribution, or natural language justification. These methods improve the readability of model behavior, but "readable" does not equal "effective." For text-based graph generation systems, truly valuable explanations need to answer at least three questions: First, can the explanation pinpoint where generation failures occurred? Second, can the explanation be verified by humans, programs, or the model? Third, can the explanation be translated into operations for the next round of generation, repair, or alignment? In other words, explanations only truly enter the aesthetic closed loop when they transform from "post-hoc explanations" into "verifiable feedback" and "executable constraints." Therefore, this chapter will explore four conditions: localizability, verifiability, executableness, and causal reliability. A detailed summary of explanation interface types can be found in Table 5.

Table 5. Explain the interface type, representative methods, and their functions and risks in generating a closed loop

Explain the interface type	**Representative method**	**Main contributions**	**Closed-loop value**	**Main risks**
Significance and Conceptual Explanation	Grad-CAM (Selvaraju RR et al., 2017); TCAV (Kim B et al., 2018)	Visualization models focus on regional or high-level concepts.	This helps in the initial diagnosis of model focus points.	Most explanations are based on correlation and are difficult to directly translate into corrective actions.
Diffusion Attribution Explanation	DAAM (Tang R et al., 2023); TNIM (Han S-H & Choi H-J., 2025)	Revealing the relationship between tokens, regions, and the denoising process.	Supports error location and generation process analysis.	Attention or attribution is not necessarily causal.
Question-and-answer	TIFA (Hu Y et al., 2023)	Transform text conditions into verifiable	Suitable for testing object, quantity, and	The VQA model's quality limits its coverage of open

explanation		questions.	semantic consistency.	aesthetic judgments.
Verbal evaluation	VIEScore (Ku M et al., 2024); MLLM critique (Huang Y, Sheng X et al., 2024; Huang Y, Yuan Q et al., 2024)	Generate natural language reasons and multidimensional evaluations.	Improve the readability of evaluations and assist human feedback.	There may be hallucinations, templated reasons, and unverifiable judgments.
Programmatic evaluation	Visual Programming (Cho J et al., 2023)	The evaluation process is broken down into executable modules.	Enhance the structure and verifiability of the evaluation process.	It relies on external vision modules, and the coverage is limited by the module's capabilities.
Regional feedback	RichHF (Liang Y et al., 2024); Focus-N-Fix (Xing X et al., 2025)	Locate the error to a region, lexical unit, or local repair target.	It can directly support masking, inpainting, or local correction.	Labeling is costly, and localized repairs may compromise overall consistency.
Iterative feedback	Divide–Evaluate–Refine (Singh J & Zheng L., 2023); RationalRewards (Wang H et al., 2026)	Use criticism, checks, or rewards for multiple rounds of correction.	The closest to a closed loop of "explanation-correction -reevaluation."	The system is highly complex and its feedback authenticity and long-term stability still need to be verified.

5.1 Locability

Because failures in text-based graph generation are often not systemic, but occur only in specific objects, attributes, spatial relationships, text tokens, or local regions, interpretations that only provide an overall evaluation will have very limited corrective effect on closed loops. Therefore, localizability is the primary requirement for effective interpretation in text-based graph systems.

At the error localization level, saliency maps and attention attribution methods can provide preliminary diagnostic clues to help identify potential problem areas and the correspondence between text tokens and image areas. For example, Grad-CAM (Selvaraju RR et al., 2017), (Tang R et al., 2023; Han S-H & Choi H-J., 2025) DAAM and TNIM provide visualization clues for the relationship between model attention areas, text tokens and image areas from the perspectives of classification response, cross-attention and diffusion denoising process, respectively. Their main contribution is to transform abstract generation failures into diagnostic information at the spatial or lexical level, enabling researchers to observe the possible correspondence between text conditions and image content. In particular, in diffusion models, (Tang R et al., 2023; Han S-H & Choi H-J., 2025) DAAM and TNIM extend the explanation object from the final image to the generation process itself, providing an entry point for analyzing the role of text conditions in the denoising trajectory.

However, it's important to note that localizability is not the same as causal explanation. Attention maps or heatmaps can indicate what the model might be focusing on, but they cannot automatically prove that the highlighted areas are the true cause of generation failure. High-response areas may only be relevant clues, not modifiable variables. Therefore, in closed-loop systems, spatial explanations are more suitable as diagnostic criteria for mislocalization and candidate correction regions, rather than being directly equated with reliable feedback. Their value lies in helping the system narrow down the scope of the problem and providing clues for subsequent validation or local correction, rather than solely making aesthetic judgments or attributing causality.

5.2 Verifiability

Valid explanations require not only localizability but also verifiability. Multimodal large language models can generate seemingly reasonable evaluation reasons, but these are not necessarily reliable. For example, pointing out that an image has "harmonious colors" or "does not reflect the requirements of the prompt." However, whether these reasons truly correspond to the image content still needs further verification. If an explanation cannot be verified, it may only be linguistically reasonable rather than factually reliable.

The question-and-answer and procedural checking methods are attempts to transform open interpretations into verifiable processes. (Hu Y et al., 2023) TIFA measures the fidelity of the generated image to its text input by checking whether the image generated by the text-generated image can answer several questions automatically generated by the visual question answering (VQA) model. (Ku M et al., 2024) VIEScore uses a multimodal large language model to evaluate the conditional image generation results and gives a reasoned natural language

judgment; its value lies in improving the readability and interpretability of the evaluation process, but these linguistic reasons themselves still need to be further tested. (Cho J et al., 2023) Visual Programming further organizes the generation and evaluation process into an executable program, so that the number of objects, layout relationships and semantic constraints can be checked modularly. Overall, these methods no longer just give an overall score, but try to transform the interpretation into question-and-answer, procedural or structured judgments, thereby improving the verifiability of the evaluation results.

However, verifiability itself is limited by the capabilities of the surrogate model. If the VQA model, detector, or multimodal large language model itself has recognition errors, the interpretation and verification process may inherit these biases. More importantly, procedural verification is generally more suitable for relatively well-defined problems such as objects, quantities, spatial relationships, and semantic consistency; for more open aesthetic judgments such as compositional aesthetics, stylistic harmony, cultural context, or artistic expression, the verification criteria are still difficult to formalize. Therefore, verifiability is an important prerequisite for interpretation to enter closed-loop feedback, but it can only improve the verifiability of the interpretation and cannot automatically guarantee the integrity of the aesthetic judgment.

5.3 Feasibility

In a closed-loop system, the value of explanation lies not only in explaining the reasons for generation failure, but also in whether it can be further transformed into an executable basis for correction. For example, prompt word rewriting, candidate image screening, region masking, local inpainting, adapter conditions, sample weighting, or training loss. Rich Human Feedback is a representative work in this direction. This study not only collects overall preferences, but also asks annotators to point out regions in the image that are inconsistent with the text, as well as parts in the text that are incorrectly expressed or omitted. Through this region-level and word-level feedback, explanation is no longer just a linguistic evaluation, but can be further transformed into operations such as data filtering, local error prediction, and mask repair. Works such as Focus-N-Fix (Xing X et al., 2025), Divide–Evaluate–Refine (Singh J & Zheng L., 2023; Wang H et al., 2026) and RationalRewards also show a similar trend: the former attempts to transform local error regions into repair targets, the latter decomposes text conditions into verifiable assertions, and iteratively improves the generation results based on error feedback; while reward models based on multidimensional criticism further explore the possibility of transforming natural language criticism into correction signals during training or testing.

These studies collectively demonstrate that interpretation is shifting from a descriptive function to an executive function. However, executive functionality also introduces new risks. First, executive feedback must be sufficiently accurate; otherwise, incorrect masks, criticisms, or question-and-answer judgments may lead the system in the wrong direction for correction. Second, local repairs do not necessarily improve overall aesthetic quality and may even disrupt stylistic consistency and compositional integrity. Third, high-quality executive feedback is typically accompanied by higher annotation and computational costs, especially at the region, word, and multi-turn feedback levels. Therefore, more granular executive interpretation is not necessarily better; it should be tailored to the specific optimization interface.

5.4 Causal Reliability

Locability, verifiability, and executableness are still insufficient to guarantee the effectiveness of an explanation. For closed-loop systems, explanations also require causal reliability, meaning that the cause they point to should have a real correlation with the generation failure, and ideally, this correlation should be verifiable through intervention. In text-based graph systems, the reliability of explanations is particularly important because once an explanation enters the optimization loop, it is no longer just descriptive text but will affect subsequent generation, model updates, or user judgments.

Attention maps, saliency maps, and multimodal language criticism may all suffer from insufficient causal reliability. Existing interpretability studies have pointed out that attention weights are not necessarily equivalent to the basis of model decisions (Jain S & Wallace BC., 2019; Adebayo J et al., 2018; Jacovi A & Goldberg Y., 2020), and saliency maps may be visually reasonable but insensitive to model parameters or data generation processes. Therefore, heatmaps or attention-highlighted areas are more suitable as relevance cues than as causal variables that lead to generation failure. Similarly, although criticism generated by multimodal large language models has strong semantic expressive power, it may be affected by cue word design, visual recognition errors, training data bias, and language template tendencies. Related reviews also show that multimodal large language models (MLLMs) may produce illusory outputs that are inconsistent with visual content (Bai Z et al., 2024; Chen Z et al., 2026), so MLLM-based criticism still needs to be tested for factual consistency and fidelity. In other words, the richness of the explanation and its reliability are not necessarily proportional. On the contrary, the more complex the explanation, the more necessary it is to further test its fidelity, stability, and operability.

Therefore, explanations in a closed loop should not be proven effective solely through examples, but their

feedback value should be tested through intervention and re-evaluation. This requirement is consistent with the call for rigorous evaluation in interpretability studies (Jacovi A & Goldberg Y., 2020; Doshi-Velez F & Kim B., 2017), and also echoes the discussion in alignment studies on agent reward over-optimization and reward hacking) (Amodei D et al., 2016; Gao L, Schulman J & Hilton J., 2023): when explanations or criticisms are converted into reward signals, the system may increase agent scores, but may not actually improve the quality of generation under human preferences. Only when explanations can withstand the test of "explanation – intervention – re-evaluation" can they truly be effective as closed-loop feedback.

5.5 Summary

Saliency maps, concept interpretations, and attention attribution can provide clues about model focus or text-image correspondences, but are generally insufficient to constitute optimization feedback on their own. Question-and-answer and procedural interpretations improve verifiability, but still need to be translated into corrective actions. Region-level feedback and iterative critique-and-refine methods are closer to a complete closed loop, but their cost, stability, and causal reliability still require further verification. Therefore, in summary, an effective interpretation should not only be readable but should also possess localizability, verifiability, executability, and reliability. In other words, only when an interpretation can locate errors, be tested, translated into corrective actions, and demonstrate its contribution in re-evaluation, does it truly become effective feedback in a closed loop.

6. Research Gaps and Future Directions

Table 6 shows the major research gaps identified in the field of text-based image generation systems and provides recommendations for future research directions, highlighting areas and aspects that have not yet been fully explored or understood in the scientific literature.

Table 6. Research Blank

Category	Reason	Future research directions
Aesthetic closed-loop research	Most existing studies only cover partial interface attempts and lack complete closed-loop verification.	Develop a unified closed-loop framework, design quantifiable and verifiable closed-loop experiments, and conduct systematic evaluations.
Utilization of high-bandwidth signals	High-bandwidth signals such as natural language criticism and saliency maps often remain at the diagnostic level and are difficult to directly translate into optimization.	Explore structuring high-bandwidth signals into executable interfaces, including combinations of local inpainting, region fine-tuning, cue word modification, and training loss.
Multi-objective and Uncertainty Optimization	The coordination of multidimensional preferences and reward signals still relies on empirical calibration and lacks a unified evaluation standard.	Construct robustness evaluation indicators for multi-objective optimization methods and study automatic weight assignment and Pareto front strategies.
Feasibility and causality of the explanation	Attention maps or natural language interpretations may lack causal reliability and have limited feasibility.	Design verifiable causal inference methods that combine explanatory outputs with intervention actions to improve the reliability of closed-loop feedback.
Personalized alignment	User preferences vary significantly, and existing methods only support a limited number of samples, leading to cold start and privacy issues.	Build a low-cost, scalable personalized alignment mechanism that supports few-shot learning, multi-round interaction, and privacy protection.
Datasets and Evaluation Criteria	The lack of fine-grained, multi-dimensional, and actionable aesthetic evaluation datasets makes it difficult to unify and compare methods.	Construct a multimodal, richly labeled benchmark dataset that includes preferences and actionable feedback, and develop a unified evaluation metric and open testing protocol.
Cross-cultural and style generalization	Research focuses on specific cultures or styles, lacking the ability to generalize aesthetic preferences.	Develop cross-cultural, multi-style, and multi-contextual preference alignment and interpretation methods to improve the broad applicability of generative systems.

7. Conclusion

This paper, through a bibliometric analysis and systematic review of computational aesthetics in text-based image generation, proposes a research framework centered on closed-loop interfaces, re-examining the role of aesthetic signals in the generation system and their transformation paths. The bibliometric analysis shows that since 2023, the number of studies on text-based image generation and computational aesthetics has grown rapidly, but research still mainly focuses on single interfaces or localized aspects, lacking systematic attempts to integrate aesthetic evaluation, preference alignment, and interpretation into a unified closed-loop system. This indicates that despite continuous innovation in technology and methods in this field, a closed-loop perspective has not yet become the mainstream research paradigm.

A comprehensive analysis of the literature reveals that while low-bandwidth signals such as scalar ratings and preference rewards are easy to optimize, they struggle to reflect the multidimensional causes of generation failures. Medium-bandwidth signals, such as rating distributions and multi-preference structures, can reveal aesthetic uncertainty, but optimization and calibration remain challenging. High-bandwidth signals, such as attribute labels, natural language criticism, and saliency maps, can provide more refined diagnostic and local correction clues, but their verifiability, feasibility, and causal reliability still need improvement. This matching problem between signal bandwidth and system interface constitutes the core bottleneck in current closed-loop optimization research.

While some studies have attempted to achieve closed-loop systems at local levels, such as converting evaluation signals into rewards for model fine-tuning or using explanatory feedback for local repair and iterative improvement, a unified theoretical framework and empirical verification have not yet been established. Existing methods still exhibit inconsistencies in signal selection, interface design, and optimization strategies, lack adaptability to cross-cultural, multi-style, and multi-user preferences, and the reliability and verifiability of high-bandwidth interpretations remain prominent issues. These factors all limit the widespread application of closed-loop systems.

Future research should focus on closed-loop optimization, systematically integrating low, medium, and high bandwidth aesthetic signals, and designing interfaces that balance optimizability, feasibility, and causal reliability. Multi-objective and uncertain optimization methods need to develop robust strategies to mitigate signal conflict and bias propagation issues. High-bandwidth interpretations should be further structured to directly support local corrections, training optimization, and user interaction. Simultaneously, personalized and interactive closed-loop strategies should consider low-sample size, privacy protection, and cold-start issues, and develop cross-cultural, multi-contextual, and multi-style datasets to establish unified evaluation criteria. Only by organically combining evaluation, alignment, and interpretation can text-based graph generation systems achieve true self-correction, continuous improvement, and robust closed-loop optimization.

In summary, despite significant technological advancements in the field of computational aesthetics for text-based graph generation, a closed-loop interface perspective remains scarce. Future research needs to systematically integrate different signal types and optimized interfaces, and design and evaluate them in a verifiable, executable, and causally reliable manner. This will enable generation systems to move from local optimization to overall closed-loop design, thereby achieving more stable, diverse, and aesthetically pleasing generation results.

References

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:160606565.

Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, Shou MZ. (2024). Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:240418930.

Chen Z, Min Y, Zhang J, Yan B, Wang J, Wang X, Shan S. (2026). A survey of multimodal hallucination evaluation and detection. *International Journal of Computer Vision, 134*, 131.

Cho J, Zala A, Bansal M. (2023). Visual programming for step-by-step text-to-image generation and evaluation. *Advances in Neural Information Processing Systems, 36*, 6048-6069.

Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Collins KM, Kim N, Bitton Y, Rieser V, Omidshafiei S, Hu Y, Chen S, Dutta S, Chang M, Lee K. (2024). Beyond Thumbs Up/Down: Untangling Challenges of Fine-Grained Feedback for Text-to-Image Generation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. vol 1. Pp. 293-303.

Dang M, Singh A, Zhou L, Ermon S, Song J. (2025). Personalized preference fine-tuning of diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 8020-8030.

Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of business research, 133*, 285-296.

Doshi-Velez F, Kim B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:170208608.

Ethayarajh K, Xu W, Muennighoff N, Jurafsky D, Kiela D. (2024). Kto: Model alignment as prospect theoretic optimization. arXiv preprint arXiv:240201306.

Feng S, Ma J, de Sa VR. (2025). FERGI: Automatic Scoring of User Preferences for Text-to-Image Generation from Spontaneous Facial Expression Reaction. In: 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, pp. 1-11.

Gao L, Schulman J, Hilton J. (2023). Scaling laws for reward model overoptimization. In: International Conference on Machine Learning. *PMLR*, pp. 10835-10866.

Ghosh D, Hajishirzi H, Schmidt L. (2023). Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems, 36*, 52132-52152.

Han S-H, Choi H-J. (2025). TNIM: A Unified Text-to-Noise Influence Mapping Approach for Diffusion Models. In: 2025 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, pp. 117-119.

Hu B, Li N, He L, Lu W, Li L, Gao X. (2025). AGIAA-2K: A Fine-grained Dataset for Aesthetic and Alignment Evaluation of AI-Generated Images. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 1-5.

Hu S. (2026). CAPO: Causal-Adaptive Preference Optimization for Diffusion Models via Causal Inference Uncertainty. IEEE Access.

Hu Y, Liu B, Kasai J, Wang Y, Ostendorf M, Krishna R, Smith NA. (2023). Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20406-20417.

Huang Y, Sheng X, Yang Z, Yuan Q, Duan Z, Chen P, Li L, Lin W, Shi G. (2024). AesExpert: Towards Multi-modality Foundation Model for Image Aesthetics Perception. Paper presented at the Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne VIC, Australia.

Huang Y, Yuan Q, Sheng X, Yang Z, Wu H, Chen P, Yang Y, Li L, Lin W. (2024). Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. arXiv preprint arXiv:240108276.

Jacovi A, Goldberg Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 4198-4205.

Jain S, Wallace BC. (2019). Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 3543-3556.

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning, PMLR, pp. 2668-2677.

Kirstain Y, Polyak A, Singer U, Matiana S, Penna J, Levy O. (2023). Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems, 36*, 36652-36663.

Ku M, Jiang D, Wei C, Yue X, Chen W. (2024). Viescore: Towards explainable metrics for conditional image synthesis evaluation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12268-12290.

Lee J, Choi J. (2026). Flow-Multi: A Flow-Matching Multi-Reward Framework for Text-to-Image Generation. *Sensors, 26*, 1120

Lee K, Li X, Wang Q, He J, Ke J, Yang M-H, Essa I, Shin J, Yang F, Li Y. (2025). Calibrated multi-preference optimization for aligning diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 18465-18475.

Lee SH, Li Y, Ke J, Yoo I, Zhang H, Yu J, Wang Q, Deng F, Entis G, He J. (2024). Parrot: Pareto-optimal

multi-reward reinforcement learning framework for text-to-image generation. In: European Conference on Computer Vision. Springer, pp. 462-478.

Liang Y, He J, Li G, Li P, Klimovskiy A, Carolan N, Sun J, Pont-Tuset J, Young S, Yang F. (2024). Rich human feedback for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19401-19411.

Lin X, Zhong S, Liu Y, Chen G. (2024). Sal-Guide Diffusion: Saliency Maps Guide Emotional Image Generation through Adapter. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp. 1-6.

Murray N, Marchesotti L, Perronnin F. (2012). AVA: A large-scale database for aesthetic visual analysis. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp. 2408-2415.

Nabati O, Tennenholtz G, Hsu C, Ryu M, Ramachandran D, Chow Y, Li X, Boutilier C. (2024). Preference Adaptive and Sequential Text-to-Image Generation. arXiv preprint arXiv:241210419.

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp. 618-626.

Singh J, Zheng L. (2023). Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative VQA feedback. *Advances in Neural Information Processing Systems, 36*, 70799-70811.

Talebi H, Milanfar P. (2018). NIMA: Neural image assessment. *IEEE transactions on image processing, 27*, 3998-4011.

Tang R, Liu L, Pandey A, Jiang Z, Yang G, Kumar K, Stenetorp P, Lin J, Türe F. (2023). What the daam: Interpreting stable diffusion using cross attention. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5644-5659.

Tong Y, Zhang J, Wei S, Guo W, Zhuang F, Wang D, Yang X, Xuan R. (2025). MindScore: quantifying human preference for text-to-image generation through multi-view lens. *Science China Information Sciences, 68*, 160105.

Wallace B, Dang M, Rafailov R, Zhou L, Lou A, Purushwalkam S, Ermon S, Xiong C, Joty S, Naik N. (2024). Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8228-8238.

Wan Y, Xiao L, Wu X, Yang J, He L. (2024). Imaginique Expressions: Tailoring Personalized Short-Text-to-Image Generation Through Aesthetic Assessment and Human Insights. *Symmetry, 16*, 1608.

Wang H, Wei C, Ren W, Liu J, Lin F, Chen W. (2026). RationalRewards: Reasoning Rewards Scale Visual Generation Both Training and Test Time. arXiv preprint arXiv:260411626.

Wu X, Sun K, Zhu F, Zhao R, Li H. (2023). Human preference score: Better aligning text-to-image models with human preference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2096-2105.

Xing X, Saha A, He J, Hao S, Vicol P, Ryu M, Li G, Singla S, Young S, Li Y. (2025). Focus-n-fix: Region-aware fine-tuning for text-to-image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 18486-18496.

Xu J, Liu X, Wu Y, Tong Y, Li Q, Ding M, Tang J, Dong Y. (2023). Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems, 36*, 15903-15935.

Zhang S, Wang B, Wu J, Li Y, Gao T, Zhang D, Wang Z. (2024). Learning multi-dimensional human preference for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8018-8027.

Zupic I, Čater T. (2015). Bibliometric methods in management and organization. *Organizational research methods, 18*, 429-472.

Paradigm Academic Press
Art and Society
ISSN 2709-9830
JUN. 2026 VOL.5, NO.2

Worship Music: Its History, Distinction from CCM, and Contemporary Challenges

Xinyi Li[1]

[1] Kyung Hee University, Gyeonggi-do 17104, Republic of Korea

Correspondence: Xinyi Li, Kyung Hee University, Gyeonggi-do 17104, Republic of Korea.

doi:10.63593/AS.2709-9830.2026.06.006

Abstract

Worship music occupies a significant position within contemporary Christian music as a distinct musical form dedicated to liturgical services. Unlike Contemporary Christian Music (CCM), which embraces diverse popular music styles for broader cultural engagement, worship music is primarily designed for congregational participation within church settings, fostering an atmosphere of sacred encounter and communal devotion. This paper provides a comprehensive historical examination of worship music, tracing its development from early Christian liturgical practices through the medieval period, the Reformation, and into the modern era. Drawing on historical and musicological scholarship, the paper analyzes the key distinctions between worship music and CCM in terms of musical style, lyrical content, performance context, and target audience. Furthermore, it examines contemporary challenges facing both genres, including commercialization, theological shallowness, and the tension between maintaining religious sacredness and achieving popular accessibility. The paper argues that while worship music and CCM share a common foundation in Christian faith, their differences in form, function, and audience reflect broader tensions within contemporary religious music between tradition and innovation, sacred and secular, and communal worship and individual consumption. Addressing these challenges requires intentional efforts to enrich musical and lyrical depth while preserving the spiritual integrity of worship practices.

Keywords: worship music, Contemporary Christian Music (CCM), liturgical music, reformation, congregational participation, commercialization

1. Introduction

Worship music is specifically designed to serve liturgical services within the church, enabling believers to immerse themselves in a sacred environment through lyrics, melody, and atmosphere, and to experience a personal connection with God (Abraham, I., 2023). Through congregational singing, worship music unites believers in collective self-reflection and spiritual elevation (Ingalls, M. M., 2018).

CCM, by contrast, represents a more diverse musical form incorporating pop, rock, and R&B. Extending beyond traditional religious ritual, CCM communicates with broader audiences through stylistically diverse music, becoming accessible church music in everyday life (Abraham, I., 2023). While transmitting core Christian values, CCM also possesses entertainment value and popular appeal, attracting younger generations and non-church audiences. Through contemporary and popular approaches, CCM has expanded the global influence of Christian music.

This paper provides a comprehensive introduction to worship music and analyzes its differences from CCM in lyrical content, usage contexts, and performance formats, exploring how they influence listeners' spiritual lives. Additionally, it examines challenges facing both genres—such as balancing religious sacredness with popular accessibility—and how to better accommodate diverse cultures in the era of globalization.

2. The Historical Development of Worship Music

Worship music, also known as praise and worship music, occupies an important position in religious music as a musical form expressing worship and praise to God, and serves as a bond connecting believers with God. Its development spans several centuries, undergoing evolution and innovation across multiple historical periods. It not only maintains deep connections with Christian faith but also exerts profound influence on a global scale.

2.1 Origins: Early Christianity to the Medieval Period

The earliest origins of worship music can be traced back to the early Christian period from the first to the fifth centuries. During this time, Christianity faced persecution under the Roman Empire, and Christian activities were often conducted underground or in secret locations. As a result, early Christian worship activities largely lacked formal music, and worship music primarily existed as part of choral singing and personal prayer, relying on the reading of Scripture, prayer, and the singing of psalms. The Psalms of Jewish worship, expressed through singing and chanting, are referenced in the New Testament with numerous exhortations for believers to praise God through songs and music—for example, Ephesians 5:19 encourages believers to speak to one another "with psalms, hymns, and songs from the Spirit."

Entering the medieval period (fifth to fifteenth centuries), following the establishment of the Christian church as the dominant force in Western European society, music gradually became more systematized. This was a critical period for the development of church worship music. As ecclesiastical authority increasingly concentrated in the hands of the Roman Pope and bishops, music became an indispensable part of church ritual, serving the celebration of the Mass and the sacraments. The primary musical form of this period was Gregorian chant, one of the most influential musical forms of the Middle Ages. Characterized by solemnity and simplicity, Gregorian chant is a monophonic, unaccompanied Latin chant typically used in the Mass and other liturgical services. It employs plainchant-smooth, flowing melodies sung in a steady manner-embodying the seriousness and piety of religious ritual. Gregorian chant features free rhythm, following the natural melody of the text rather than a fixed meter (Wilson-Dickson, A., 1992).

2.2 The Reformation and the Democratization of Worship Music

The next significant period was the Renaissance and Baroque era, following the Protestant Reformation. Martin Luther proposed that "believers can communicate directly with God through singing" and promoted the use of hymns. He firmly believed that all believers could praise God in their own language; consequently, German hymns gradually replaced Latin chants, initiating a wave of localization of church language (Wilson-Dickson, A., 1992).

One of Luther's most important contributions to the future development of worship music was his insistence that worship should no longer be the exclusive domain of clergy but should become an activity involving all believers. This directly influenced the usage contexts and forms of worship music. For worship music, this meant it was to be a music of the people—sung collectively by congregations (Ingalls, M. M., 2018). Luther's reforms diversified the formerly uniform performance format of church music.

Luther composed the classic hymn *A Mighty Fortress Is Our God* (1527–1529), drawing on Psalm 46. Examining the musical score reveals the simplicity of its chords, rhythm, and melodic progression. By localizing church music language and using simple, easily singable melodies, Luther directly influenced the development of worship music, making church music more accessible to ordinary people (Wilson-Dickson, A., 1992). (see Figure 1)

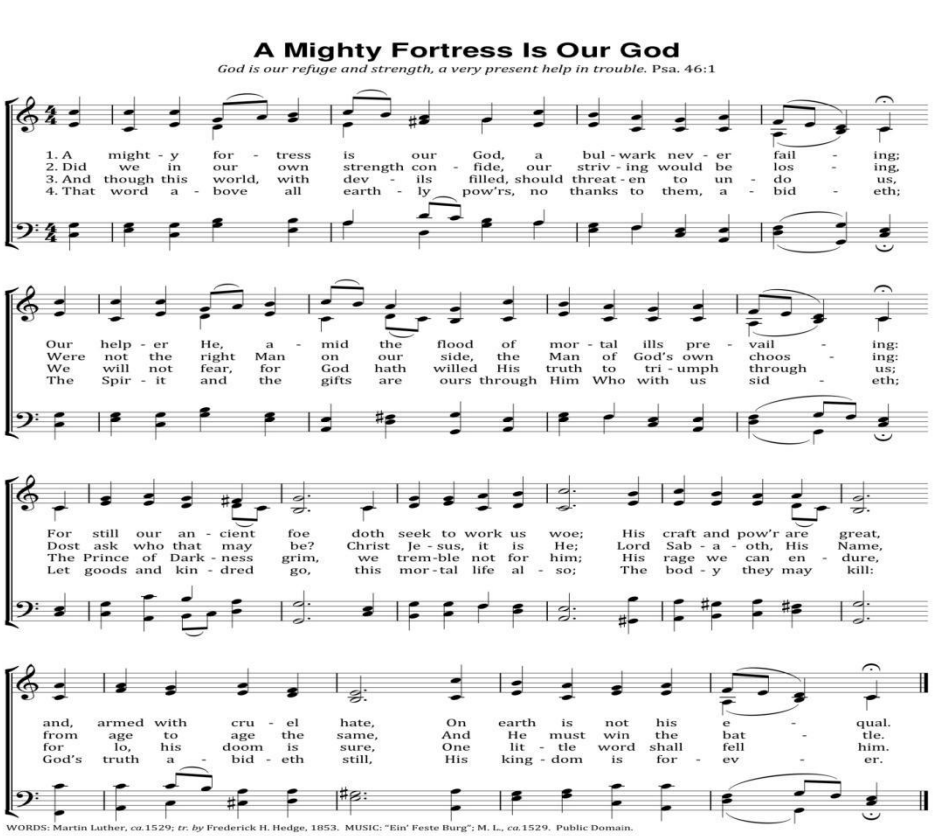

Figure 1. Musical score of Martin Luther's hymn *A Mighty Fortress Is Our God*

2.3 The Enlightenment to the Modern Era

Entering the eighteenth and nineteenth centuries, the style of worship music underwent significant changes. With the arrival of the Enlightenment, musical development became increasingly diverse. The Enlightenment emphasized reason, individual freedom, and universal education, which likewise led to participation in worship music no longer being limited to clergy or professional musicians, but opening up to a broader community of believers. Musical structures were no longer as complex as in the Baroque period; instead, they became simpler, easier to teach, and more readily disseminated (Payne, L., 2024). This secularization of music meant that more people could participate in church worship choir groups, and activities became more varied, including family Bible studies, women's or working professionals' gatherings, and more.

3. Lyrical Content of Worship Music

The lyrical content of worship music includes praise, adoration, and thanksgiving to God, centering on the redemption of Jesus Christ and the attributes of God. It narrates God's greatness and glory, His love and grace, the redemption and sacrifice of Jesus Christ, the believer's dependence and trust, God's holiness and justice, and corporate worship and shared faith (Ingalls, M. M., 2018).

Taking the song *How Great Is Our God* as an example, the lyrical content primarily praises God's greatness. The central theme — "how great is our God" — is repeated throughout the song. The lyric "He wraps himself in light" uses "light" as a symbol of God's purity and holiness, representing the manifestation of God's glory. "And darkness tries to hide" suggests that God's light cannot be obscured by darkness, reflecting God's omnipotence and omnipresence (Ng, S., 2022). The lyrics of this song extol God in a simple yet powerful manner, allowing listeners to immediately grasp the central message. By emphasizing God's greatness in the life of the universe, the song guides believers to experience God's majesty and holiness in worship, inspiring infinite reverence and adoration. (see Figure 2)

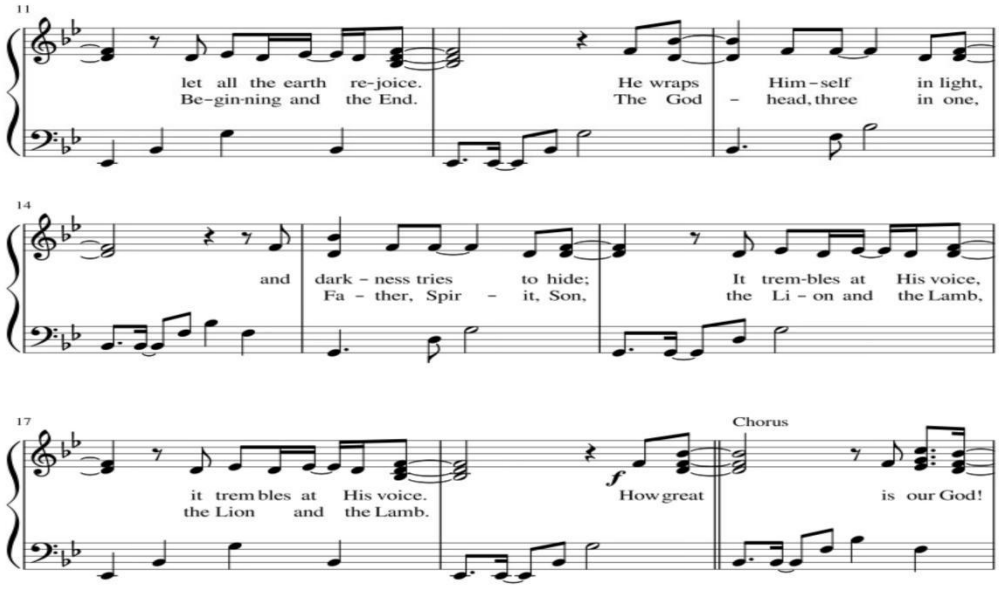

Figure 2. Example of lyrical structure in contemporary worship music (e.g., "How Great Is Our God")

4. Distinctions Between Worship Music and CCM

Having analyzed the lyrical content of worship music, we can further discuss the distinctions between worship music and CCM. Although both are part of Christian music and centered on God, their styles, purposes, and modes of expression differ. Worship music focuses more on helping believers draw near to God in collective worship and express their reverence, while CCM typically incorporates more popular music elements to reach a broader audience (Abraham, I., 2023). I identify three primary areas of difference: musical style, usage context, and target audience.

4.1 Musical Style

In terms of musical style, worship music features relatively simple melodies, harmonies, and rhythms, designed to enable congregational participation and create a shared worship experience. A steady, even rhythm, simple and memorable harmonic progressions, and accessible arrangements constitute the primary stylistic characteristics of worship music. For CCM, however, musical structures are more complex, incorporating more contemporary musical elements with richer and more varied melodic contours, rhythmic grooves, and arrangement structures (Abraham, I., 2023). CCM incorporates more entertainment-oriented elements that align with contemporary trends and popular tastes. Representative artists and groups include MercyMe and Newsboys (Payne, L., 2024).

Regarding lyrical content, CCM is no longer limited to worship of God but also includes reflections on life's challenges and uncertainties, incorporating more personal emotion. For example, the song Cinderella by Steven Curtis Chapman expresses a father's genuine love for his daughter and his reluctance to see her grow up, conveying complex emotions of joy, nostalgia, and letting go (Payne, L., 2024). This song resonates with the emotions of many families and has become a classic CCM song. The ability to find relatable emotions within various personal experiences and to feel and share them has become a hallmark of CCM. Compared to worship of God, such lyrical content and emotion are more secularized, closer to everyday life, and easier to empathize with—which is precisely why CCM has secured a place in the popular music landscape today.

4.2 Usage Context and Target Audience

In terms of usage context and target audience, worship music is primarily used in church worship services, often performed as congregational singing in group settings (Ingalls, M. M., 2018). The emphasis is on involving every believer in the activity, uniting them in shared experience. CCM, by contrast, is used in more diverse contexts beyond the church, including personal listening, radio broadcasts, concerts, and other formats and venues. As a form of culture and entertainment, CCM aims to convey values and touch personal emotions. Although the themes of CCM songs remain related to Christian faith, they possess greater commercial and popular appeal, and can be heard in cars, at family gatherings, in cafes, and elsewhere.

In terms of target audience, worship music primarily reaches Christian believers, especially those participating in worship services, who often have high expectations for participatory singing. For CCM, the audience is not limited to Christians but also includes many non-Christians who are drawn to themes of faith and positive

energy. Because CCM songs often address emotions in everyday life — love, family, life's struggles, and more — they are particularly popular among contemporary youth, especially those facing life's difficulties and challenges, who can find spiritual comfort in the music (Fiske, J., 1989).

5. Contemporary Challenges and Future Directions

As worship music has developed globally, problems such as excessive commercialization, stylistic homogeneity, and lack of lyrical depth have gradually appeared.

First, lyrical content often lacks depth. Many worship songs express simple, repetitive sentiments — "I love You, Lord," "You are our only One" — while neglecting the complex and profound intellectual content of religious theology and its foundational doctrines. An excessive focus on personal emotional experience overlooks broader social issues and the church's missional calling (Sigmon, C. T., 2022).

Second, there is a tendency toward excessive commercialization. Worship music is increasingly treated as a profit tool, with some music companies signing worship teams and over-packaging them, losing sight of worship music's original purpose (Payne, L., 2024).

Addressing these challenges requires innovation in both musical form and social engagement. First, musical quality should be enhanced by incorporating traditional and local elements while deepening lyrical content to balance theology and emotion. Second, social issues — such as environmental protection and public welfare — should be integrated into songs, enabling worship music and CCM to become calls to social action (Miller, D. E., 1997). Worship music can also reach society through charity concerts and religious cultural events, fostering understanding of core Christian values and expanding Christianity's global influence.

6. Conclusion

Through a comprehensive introduction to worship music and analysis of its differences from CCM, we can see that although both are essentially centered on transmitting Christian faith, they differ in expressive form, target audience, and usage context (Abraham, I., 2023). These differences not only enrich the expressive power of Christian music but also provide the public with more choices. However, in the development of both genres, issues such as commercial tendencies and musical homogenization have also emerged. For worship music and CCM to better fulfill their roles in society and the church in the future, we must think more deeply about the meaning of the music itself, innovating and adjusting while maintaining faith. Only in this way can worship music and CCM develop in a balance between globalization and secularization, making greater contributions to both the church and society.

References

Abraham, I. (2023). Contemporary Christian music and contemporary worship music. In D. Hiebert (Ed.), *The Routledge international handbook of sociology and Christianity* (pp. 242–253). Taylor and Francis.

Fiske, J. (1989). *Understanding popular culture*. Routledge.

Ingalls, M. M. (2018). *Singing the congregation: How contemporary worship music forms evangelical community*. Oxford University Press.

Miller, D. E. (1997). *Reinventing American Protestantism: Christianity in the new millennium*. University of California Press.

Ng, S. (2022). Musical eschatology in contemporary Christian worship songs. *Music Theory Online, 28*(4). https://doi.org/10.30535/mto.28.4.ng

Payne, L. (2024). *God gave rock and roll to you: A history of contemporary Christian music*. Oxford University Press.

Sigmon, C. T. (2022). Blessed is the one whose bowels can move: An essay in praise of lament in contemporary worship. *Religions, 13*(12), 1161. https://doi.org/10.3390/rel13121161

Wilson-Dickson, A. (1992). *The story of Christian music: From Gregorian chant to black gospel*. Fortress Press.

Paradigm Academic Press
Art and Society
ISSN 2709-9830
JUN. 2026 VOL.5, NO.2

From Audience to Co-Creator: Participatory Public Program Design Around Fotografiska Shanghai's The Pet Show

Pinyi Zhu[1]

[1] University of New South Wales, Chengdu, Sichuan, China

Correspondence: Pinyi Zhu, University of New South Wales, Chengdu, Sichuan, China.

doi:10.63593/AS.2709-9830.2026.06.007

Abstract

Public programs have become an important means through which cultural institutions extend exhibition experiences beyond passive viewing. Yet in practice, "participation" is often treated as a general indicator of audience engagement, with insufficient distinction between the forms of interpretive agency that participants actually acquire. Taking Fotografiska Shanghai's The Pet Show as an existing exhibition context, this article analyzes three proposed public programs conceived and designed by the author around the exhibition. Using a design-oriented curatorial case-study approach, it examines how structural design can create conditions for audiences to move from spectators to co-creators of cultural meaning, while comparing participatory mechanisms developed for different publics. Theoretically, the article brings together Nina Simon's participatory museum model, Claire Bishop's critique of participatory art, and Suzanne Lacy's account of socially engaged practice. It proposes three modes of co-creation—material, narrative, and discursive—and identifies emotional perception, creative expression, and critical reflection as three parallel entry points into participation. The article argues that the analysis of participatory curating should focus not on the number of activities offered, but on how interpretive agency is redistributed and how participant-generated content enters processes of production, negotiation, and circulation. Because the three public programs remain curatorial design proposals rather than official programs implemented by Fotografiska Shanghai, the article limits its claims to the participatory possibilities supported by the designs and does not present anticipated effects as observed audience change. The framework is intended as an analytical and design tool that can be tested through subsequent practice and empirical research.

Keywords: participatory curating, public programs, curatorial design, audience participation, co-creation, Fotografiska Shanghai

1. Introduction

Public programs are increasingly central to the ways cultural institutions interpret exhibitions, build relationships with communities, and define their public roles. Artist talks, workshops, collaborative making, and facilitated discussions can extend exhibitions temporally and socially, enabling audiences to respond to what they encounter, bring personal experience into institutional space, and produce forms of knowledge that cannot be accommodated through display alone (Bishop, 2012). This development reflects a broader shift in curatorial practice from the one-way transmission of authoritative interpretation toward more dialogic and participatory forms of cultural mediation. Within such a model, audiences are no longer merely the final recipients of curatorial information but become part of the processes through which meaning is produced, contested, revised, and circulated (Simon, 2010).

However, "participation" itself is a concept with blurred boundaries. A visitor pressing an interactive button, a child completing a drawing, a photographer telling the story behind an image, and an adult taking part in an ethical discussion may all be described as "participants," yet the degree and nature of their agency are clearly

different. For public programs, this ambiguity matters. If participation is understood primarily in terms of attendance, activity levels, or entertainment, the language of "co-creation" can remain rhetorical rather than analytically meaningful. A cultural institution may appear highly participatory while still retaining almost complete control over the questions being asked, the range of legitimate interpretations, and the outcomes presented. A more useful analysis therefore needs to distinguish between "being active" and possessing interpretive agency.

This article compares three proposed public programs developed in relation to Fotografiska Shanghai's The Pet Show. It is important to clarify that The Pet Show constitutes the pre-existing exhibition and thematic context of this study, while Pets' Feelings, Pets' Eyes, and Pets' Impact are independent curatorial designs by the author rather than official public programs published or implemented by Fotografiska Shanghai. Pets' Feelings is designed for children aged five to ten and their families, centering on collaborative parent–child drawing. Pets' Eyes targets young people aged approximately fifteen to twenty-five and centers on photography, storytelling, and peer response. Pets' Impact is intended for adults and uses interdisciplinary discussion and structured interaction to address pet companionship, well-being, and human–animal ethics. The three programs use emotional perception, creative expression, and critical reflection as their principal entry points, but these are treated as parallel design strategies rather than as age-based hierarchies of ability or participatory intensity.

The central research question is: how can exhibition-based public programs create conditions for audiences to move from passive spectators to active co-creators of cultural meaning? Two related questions follow. First, how should participatory mechanisms be differentiated for different publics? Second, how can public programs extend the social and ethical issues of an exhibition beyond the gallery display? Because the objects of analysis are curatorial designs rather than implemented interventions supported by systematic audience data, the study does not seek to measure changes in participant attitudes or behavior. Instead, it analyzes the participatory structures constructed by the three programs and the different forms and degrees of co-creative space that these structures make available to audiences.

This article argues that the key to participatory curating lies not in simply increasing the number of audience activities, but in how curatorial structures reconfigure relations of power in meaning production and interpretation. Its conceptual contribution is twofold. First, it distinguishes three principal modes of co-creation—material, narrative, and discursive—in order to identify what participants actually produce. Second, it connects these modes to the principal entry points through which different publics participate, providing a means of analyzing relationships among structure, agency, and the circulation of meaning. At the level of practice, the article further proposes an audience-responsive design logic that can help cultural institutions clarify participant roles, forms of contribution, and mechanisms of circulation.

2. Literature Review and Conceptual Framework

2.1 From Viewing to Participation

Nina Simon's The Participatory Museum provides an important foundation for understanding changing audience roles in cultural institutions. Simon (2010) argues that the core of a participatory cultural institution is not simply the addition of interactive activities, but the use of institutional and programmatic design to enable audiences to move from recipients of content to contributors who participate in the production of cultural meaning through personal experience, knowledge, objects, and social relationships. Compared with traditional institution-centered communication models, participatory practice places greater emphasis on audience agency in processes of content production and interpretation. Similarly, Black (2005) argues that contemporary museums need to move away from one-way knowledge transmission toward more responsive and participatory relationships with audiences, allowing audience experience to become part of museum meaning-making.

In this sense, participation does not require curators to abandon professional judgment; rather, it entails a partial shift in the focus of curatorial work. Curating involves not only selecting works, organizing knowledge, and designing displays, but also designing the conditions of participation: invitations, tasks, sequences of activity, interactive interfaces, and social rules among participants. This article describes this process as a participatory architecture: a curatorial structure that defines how audiences enter a program, what they can contribute, and how those contributions can be displayed, responded to, and circulated further. This is consistent with Simon's (2010) participatory design principles: effective participation does not depend on openness alone, but requires clear structures, participatory goals, and intelligible relationships between participants and institutions.

This structural understanding is particularly important for public programs. An open invitation to participate does not automatically generate meaningful contributions; participants also need to understand their roles, what they are contributing to, and how their contributions will be seen and responded to by others. Drawing on Simon (2010), this article therefore treats "participatory architecture" as a unit of analysis, focusing on role definition, task sequence, the visibility of contributions, and mechanisms for the circulation of meaning rather than equating

workshops, voting, or discussion with participation in themselves.

Simon's theory also helps distinguish accessibility from participatory agency. Free admission, weekend scheduling, lower technical thresholds, and materials suited to different age groups can all make a program more accessible, but accessibility alone does not constitute co-creation. Hooper-Greenhill's (2000) research on museum learning demonstrates that audiences do not passively receive predetermined knowledge; they actively construct meaning through personal experience, prior knowledge, and social context. Yet audience interpretation can become agency within participatory curating only when institutional structures provide spaces for that meaning to be expressed and circulated. Co-creation therefore implies that at least some responsibility for content production or interpretation shifts from the institution to participants. A program with large attendance may remain highly institution-led, while a smaller program may generate intensive collaborative meaning-making (Lynch, 2011).

2.2 From Interaction to Co-Creation

Claire Bishop's critique of participatory art provides an important theoretical basis for distinguishing interaction, participation, and co-creation. Bishop (2012) rejects the assumption that participation is inherently democratic, emancipatory, or socially progressive, arguing instead that participatory art must be evaluated in relation to its specific organization, power structures, aesthetic experiences, and social relations. In other words, the existence of participation is not in itself evidence of value; what matters is the degree of action and interpretive space made available to participants.

This argument is particularly relevant to public programs in cultural institutions. A workshop may contain numerous physical activities and interactive elements while leaving core interpretive authority in the hands of the curator or facilitator. Conversely, a discussion that produces no material object may grant participants considerable interpretive agency through sustained exchange and negotiation of meaning. This article therefore does not use the number of activities or frequency of interaction as its primary measure of participation. Instead, it asks whether audiences actually enter processes of meaning production, interpretation, and negotiation. This distinction also resonates with Carpentier's (2011) discussion of participation: participation differs from contact or interaction more generally because it concerns whether participants acquire a substantive position in decision-making, meaning formation, or the distribution of power.

On this basis, the article makes an analytical distinction among interaction, participation, and co-creation. Interaction primarily refers to participants responding to content or tasks predetermined by the institution. Participation means that participants' actions, choices, or contributions have some influence on the program process. Co-creation goes further, referring to participants entering processes of content production, interpretation, negotiation, and the circulation of meaning. These concepts do not necessarily form a simple hierarchy of value. Interaction can itself be valuable for guided tours, information access, or lowering barriers to engagement, and not every component of a public program needs to pursue intensive co-creation. The key question is whether a program contains identifiable mechanisms that redistribute interpretive agency so that meaning production is no longer completed entirely by the institution (Bishop, 2012).

This conceptual distinction also reveals that co-creation can take different forms. The first is material co-creation, in which participants make images, texts, objects, or display content that becomes part of the public program. The second is narrative co-creation, in which stories, captions, oral experiences, and peer responses alter how an object or theme is understood. The third is discursive co-creation, in which participants enter public discussion and negotiate among different forms of knowledge, experience, and value positions. Museum research has likewise shown that the significance of participatory practice lies not only in content production but also in the renegotiation of knowledge relations, institutional authority, and public voice (Lynch, 2011). Distinguishing among different modes of co-creation therefore avoids reducing participation to hands-on making and makes different types of public program analytically comparable.

2.3 Public Programs as Social Practice

Lacy (1995) argues that artistic practice can move beyond the production of conventional objects and become a social practice by organizing relationships, generating dialogue, and gathering social issues. Within this framework, artists and curators are not only producers of objects but may also become organizers of social relations and facilitators of public discussion. Public programs in cultural institutions can therefore be understood not merely as educational supplements to exhibitions, but as temporary forms of social infrastructure through which participants with different forms of knowledge and lived experience gather around shared issues.

This approach resonates with Kester's (2004) work on dialogical art practice. Kester emphasizes that dialogue itself can constitute an important component of artistic and cultural production; its value need not be embodied in a final material object, but may lie in ongoing processes of exchange, negotiation, and understanding among different subjects. Thus, when public programs address issues such as animal ethics, care, mental health, or

social responsibility, discussion and relationships can themselves become important sites of curatorial meaning production.

This relational perspective also provides a theoretical basis for the distinction made later in this article between "audiences" and "publics." Warner (2002) understands a public as a relational collectivity formed through the circulation of and response to discourse, rather than as a group defined in advance by demographic characteristics. For public programs, this means that a "public" is not only identified before an event begins; it may also be temporarily constituted as participants view, respond to, and negotiate one another's contributions.

In Empathic Vision, Bennett (2005) proposes that empathic viewing is not a simple projection of the viewer's own experience onto another, but a form of affective encounter through visual and sensory experience that simultaneously preserves the difference of the object. For this study, the value of the concept is not that it presupposes participation will "increase empathy," but that it helps analyze how emotional perception can be designed curatorially as an entry point into observation, expression, and reflection.

Azoulay (2008) understands photography as an ongoing social relation among photographer, photographed subject, and viewer rather than as a neutral visual record. The meaning of a photograph therefore continues to be generated through viewing, circulation, response, and reinterpretation. For the youth photography program discussed here, this perspective provides an analytical lens: participants not only produce images but also confront how their representational choices organize relations of looking between humans and animals.

Taken together, Simon, Bishop, and Lacy form the principal theoretical chain of this article. Simon (2010) explains how cultural institutions can create conditions for audience contribution through structural design; Bishop (2012) provides a critical criterion for judging whether participation actually produces interpretive agency; and Lacy (1995) situates participation within the production of public dialogue and social relations. Bennett (2005) and Azoulay (2008) add analytical dimensions concerning affective viewing and the relationality of photography, while Warner (2002) supports the later discussion of how publics are formed through interaction. On this basis, the article understands participatory public programming as a curatorial practice that reconfigures relationships among content production, interpretation, and the circulation of meaning.

2.4 Analytical Framework: Three Modes of Co-Creation

Based on this literature, the article constructs two interconnected analytical dimensions. The first concerns the principal entry point through which a public enters the program: emotional perception, creative expression, or critical reflection. These are neither fixed psychological stages nor indicators of ability across children, young people, and adults; rather, they summarize the most prominent curatorial invitation offered by each program. The second dimension concerns the dominant form of co-creation that emerges once participants enter processes of meaning production and is used to compare how different programs redistribute opportunities for content production, interpretation, and response.

The second dimension conceptualizes co-creation as material, narrative, and discursive. Material co-creation concerns how participant-generated works or visual content enter public presentation; narrative co-creation concerns how personal experience, stories, and peer interpretation alter the ways objects are understood; and discursive co-creation concerns how different forms of knowledge, experience, and value positions become shared issues through discussion and response. The three modes can overlap within a single program. Their purpose is not to rank the intensity of participation, but to identify the particular process of meaning production into which participant contributions enter.

Figure 1 summarizes this framework. The three pathways are presented in parallel in order to avoid interpreting emotional perception, creative expression, and critical reflection as an age-based progression. The analysis instead asks how different curatorial structures connect particular participatory entry points with modes of co-creation and, on that basis, compares relationships among audience characteristics, participatory mechanisms, and anticipated cultural outcomes.

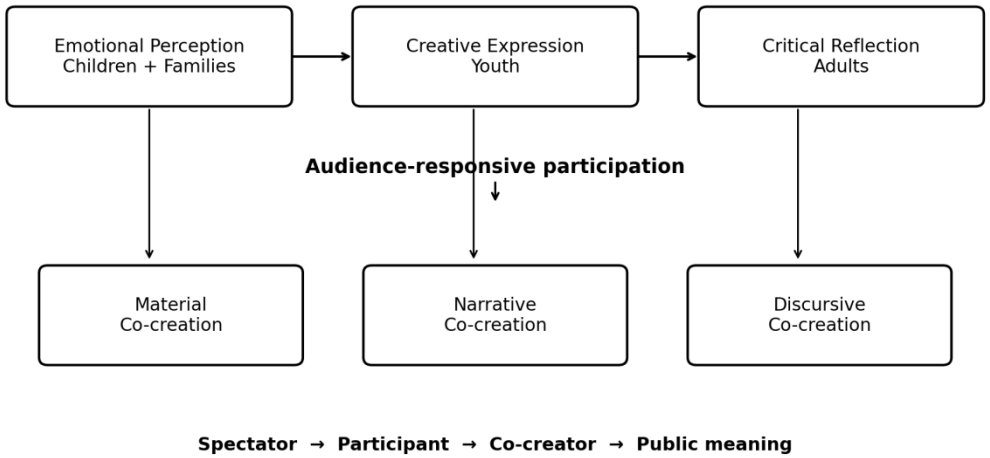

Figure 1. Audience-responsive participatory curating framework proposed in this study

3. Research Method and Case Context

3.1 A Design-Oriented Curatorial Case Study

This article adopts a design-oriented curatorial case-study method, using a set of public program proposals that have not yet been implemented as its analytical material. Unlike practice-based research focused on the outcomes of implemented projects, this study examines how participation is pre-structured within curatorial proposals: what roles are assigned to audiences, how activities are sequenced, which contributions can enter public presentation, and which interpretive and ethical responsibilities remain with the institution. The three programs are therefore treated as comparable design propositions rather than interventions whose effects have already been validated.

The analytical materials consist of complete public program proposals, including program descriptions, audience segmentation, activity sequences, facilitation structures, anticipated outputs, and theoretical rationales. To ensure consistency across cases, the study uses a four-dimensional analytical matrix: first, through what entry point does the target public encounter the exhibition theme; second, what are participants invited to produce or contribute; third, what opportunities for interpretation and response does the design provide; and fourth, how are these contributions displayed, exchanged, or circulated further? This matrix is used to identify shared design principles across the three programs and differences in the distribution of agency.

The researcher also participated in the curatorial design of all three public program proposals. This position enables design rationales and decision-making processes to be incorporated directly into the analysis, but it also introduces the risk of confirmation bias. To mitigate this risk, the article does not treat the curator's subjective intentions as evidence of effects. Instead, it analyzes the structural conditions of the three programs using the theoretical framework and consistent comparative dimensions outlined above, reserving all claims about actual audience responses for future empirical research.

Because the objects of study are design proposals, the article limits its claims to the participatory possibilities supported by the designs and makes no causal inferences about actual participatory outcomes. Expressions such as "support," "invite," and "create conditions for" are used to describe the roles and spaces for action made available by curatorial structures. Claims such as "increasing empathy," "changing attitudes," or "producing behavioral change" would require post-implementation questionnaires, interviews, observations, or other empirical evidence.

3.2 Case Context: The Pet Show and Fotografiska Shanghai

This article takes Fotografiska Shanghai's photography exhibition The Pet Show as its curatorial context. Using photography as its principal medium, the exhibition addresses human–animal relationships as emotional, social, and cultural questions. Fotografiska's institutional philosophy emphasizes the role of photography in raising public awareness and fostering social dialogue; public programming can therefore function as a curatorial mechanism for extending exhibition content beyond visual display (Fotografiska, n.d.). Within this context, the present project does not treat companion animals merely as objects to be viewed, loved, or consumed, but as cultural subjects through which family relationships, social identity, care, control, and ethical responsibility can be discussed.

Two artistic practices in the exhibition directly informed the public program designs. Wang Hanlin's Animal Capital situates animals within contemporary urban life and human social structures, offering the family workshop a visual entry point for discussing family relations, social structures, and consumer culture (Fotografiska Shanghai, n.d.-b). Sophie Gamand's photographic practice, particularly the Wet Dog series, uses animal expression, bodily states, and vulnerability to expose tensions among human control, emotional projection, and animal subjectivity, providing young participants with a critical reference point for reconsidering conventional "cute" representations of pets (Fotografiska Shanghai, n.d.-a).

At the level of program design, the three proposals employ an operational form of audience segmentation. The family workshop is primarily intended for children aged five to ten and their parents; the youth photography program is aimed mainly at participants aged approximately fifteen to twenty-five; and the adult program primarily targets participants around twenty-five to forty-five years old. These age ranges are used to organize different participatory mechanisms rather than to essentialize the capacities of particular groups. Table 1 presents only the principal entry points and design intentions of the three programs; their specific differences are analyzed in Sections 4 and 5.

Table 1. Comparative overview of the three proposed public programs

Program / Public	Primary Entry Point	Participation Mechanism	**Dominant co-created outcome**
Pets' Feelings Children + families	Emotional Perception	Exhibition looking, parent-child drawing, sharing, display	Material artifact + shared interpretation
Pets' Eyes Youth	Creative Expression	Photography, storytelling, peer response, voting	Narrative meaning + peer interpretation
Pets' Impact Adults	Critical Reflection	Expert framing, facilitated interaction, questions, ethical dialogue	Discursive meaning + public reflection

4. Participatory Design for Three Publics

4.1 Pets' Feelings: Material Co-Creation Through Family Collaboration

Pets' Feelings is designed for children aged five to ten and their parents and consists primarily of exhibition viewing, collaborative parent–child drawing, sharing, and on-site display. The key to the project is not the addition of drawing as an interactive activity, but the placement of making after exhibition viewing so that visual creation becomes a means through which participants respond to exhibition content. Participants translate their observations of animals' expressions, behaviors, and social relationships into a jointly produced visual work. Drawing is therefore not a craft activity detached from the exhibition, but a form of visual interpretation.

The program begins with facilitated looking. Unlike a conventional tour centered on knowledge transmission, this stage does not prescribe a single correct interpretation. Instead, questions concerning animal expression, posture, emotion, and relationships direct attention while enabling participants to form their own observations and judgments. For children, emotion and behavior provide accessible cues for interpreting images and allow personal observation to become the basis for subsequent making. Parents, meanwhile, move beyond the roles of chaperone and supervisor to participate with children in looking, discussion, and negotiation of meaning. The family thus becomes the basic participatory unit of the program.

During collaborative drawing, interpretation is further transformed into content production. Children and parents jointly decide what to depict, which situations and emotions matter, and what narrative emphasis the work should carry. Rather than asking participants to reproduce existing works from the exhibition, the program allows them to generate new visual content based on their own interpretations. In curatorial terms, this means that part of the production of meaning shifts from institutionally predetermined exhibition content to participating families. The drawing can therefore be understood as a material record of participants' reinterpretation of the exhibition theme.

Sharing and temporary display further strengthen this mechanism. If drawing remained within the family unit, participation would largely remain a private form of collaboration. When families introduce their works to others and place them in a public display area, audience-generated content acquires a new public status. The cultural institution continues to determine the theme, time, space, and basic rules of the program, but the visible curatorial content is no longer produced entirely by the institution. Audience interpretations gain an opportunity to enter public cultural space, echoing Simon's (2010) account of the shift from content recipient to contributor.

This article describes this mode as material co-creation. Co-creation here is not primarily judged by the artistic quality of the resulting work, but by the processes of shared observation, family negotiation, and visual interpretation that the work carries. The material outcome is a visible vehicle for the participatory process; its curatorial significance derives from the continuous relationship among looking, making, and public sharing. This also demonstrates that hands-on participation alone is insufficient to constitute co-creation. Participant agency becomes more clearly positioned only when making is connected to exhibition interpretation and public expression.

At the level of accessibility, weekend scheduling, small-scale organization, and child-friendly materials are intended to lower practical barriers to participation. These conditions, however, are only the basis for entering the program. More directly related to co-creation is the positioning of the family as a collaborative unit within processes of looking, negotiation, and expression, allowing children and parents jointly to assume part of the responsibility for interpretation and content production.

4.2 Pets' Eyes: Narrative Co-Creation Through Photography and Peer Interpretation

Pets' Eyes is designed for young participants aged approximately fifteen to twenty-five. Photography serves as the primary medium, while image production, storytelling, on-site display, and peer response form the participatory structure. Compared with the family workshop, this program gives participants a higher degree of autonomy over content: they decide what images to make or select, organize their own interpretations, and view, respond to, and evaluate the work of others.

Sophie Gamand's photographic practice functions not simply as a technical model but as a critical entry point for reconsidering the visual culture of pets. In everyday social media and consumer culture, companion animals are frequently represented through familiar visual conventions of cuteness, humor, intimacy, and companionship. Although such images are highly shareable, they may also simplify animals' complex emotional and social relations. By introducing Gamand's attention to animal expression, bodily vulnerability, human control, and emotional projection, the program reframes photography from everyday documentation as a representational practice open to critical reflection, encouraging participants to consider how images construct relationships between humans and animals.

Within this structure, the photograph is only the first layer of participant contribution. The second is narrative: participants explain the situation, emotion, relationship, or personal experience behind the image. Image production is thus connected to verbal interpretation, and participants become not only photographers but narrators of the meanings of their own works. Through this mechanism, interpretive authority within the exhibition and public program is no longer located exclusively with artists, curators, or facilitators, but is partially opened to participants themselves. Personal narratives also allow context, affect, and uncertainty that cannot be directly represented in the image to enter public exchange.

The third layer of participation comes from peer interpretation. Participants not only display their own photographs but also view, listen to, and respond to the work of others. Voting is not intended to establish an objective ranking of artistic quality; rather, it operates as a social mechanism that gives participants a reason to look carefully at and compare different narratives. Each participant therefore occupies multiple roles—photographer, narrator, viewer, and respondent. The meaning of an image is no longer determined by a single subject but continues to emerge through the creator's account, peers' viewing, and collective response.

This article describes this mode as narrative co-creation. Its principal outcome is not a collection of independent photographs but a temporary network of meaning constituted by images, personal narratives, and peer responses. In relation to Simon's (2010) discussion of audience contribution, the program creates structural conditions for participants to produce and interpret content. Bishop's (2012) critique of participatory practice, meanwhile, reminds us that meaningful participation cannot be judged solely by activity format; it requires attention to the actual degree of expressive and interpretive space participants receive. By this criterion, young participants do more than complete predetermined tasks: to some extent, they participate in the formation and circulation of the program's meaning.

Azoulay's (2008) account of the relationality of photography further reveals the ethical dimension of this design. Photography does not simply record a pre-existing object; it also establishes relations among photographer, photographed subject, and viewer. In companion-animal photography, participants can therefore consider not only whether an image is visually appealing but also how they look at, select, and represent animals. The program cannot demonstrate that this design will necessarily change attitudes or increase ethical awareness, but it can make previously implicit relations of representation available for discussion and response. In this way, the public program extends the exhibition's concern with animal representation and places part of the responsibility for interpretation with participants.

Compared with the family program, the youth project also expands the social scope within which co-creation

occurs. Whereas the former primarily uses the family as its collaborative unit, Pets' Eyes uses peer viewing and response to turn the wider participant group into a temporary interpretive community. Co-creation therefore does not only mean audiences providing content to an institution; it can also involve participants collectively producing social conditions for viewing and exchange, enabling individual experiences to be reinterpreted and responded to by others.

4.3 Pets' Impact: Discursive Co-Creation Through Dialogue and Ethical Reflection

Pets' Impact is designed for adult audiences and is organized around interdisciplinary knowledge sharing, structured interaction supported by an animal-welfare organization, and open discussion. Proposed elements include a psychologist's introduction to pet companionship and well-being, knowledge concerning animal behavior and welfare, professionally supported interaction with animals, and facilitated group discussion. Compared with the drawing and photography programs, this project lacks an obvious material-making component. Its participatory quality should therefore not be judged by whether it produces visible objects, but by whether audiences are able to enter processes of knowledge exchange and public discussion.

From this perspective, expert-led components are not necessarily incompatible with participatory curating. The key question is how professional knowledge enters the public program. If expert opinion is transmitted unidirectionally as an unquestionable conclusion, audiences remain relatively passive. If such knowledge is designed as a starting point for discussion and participants are able to bring their own experiences, questions, and differing positions into the exchange, expert knowledge can become a resource for shared meaning-making. The curatorial focus of this program is therefore not merely to invite professionals into the exhibition space, but to build relationships among artistic context, professional knowledge, and the lived experience of the public.

Structured interaction supported by an animal-welfare organization adds an embodied dimension to this process. Participants can encounter animals within clearly defined welfare and safety boundaries, but such contact should not itself be equated with co-creation. Its curatorial function is primarily to provide a shared experience for subsequent discussion, so that questions of companionship, care, animal agency, and human responsibility do not remain entirely abstract.

This article describes this mode as discursive co-creation. Unlike material co-creation, its principal output is not a physical object; unlike narrative co-creation, it does not center primarily on individual stories. Instead, it focuses on reciprocal responses among different forms of knowledge, experience, and ethical position. The institution frames the topic, invites professional contributors, and establishes the structure for discussion, but the meaning that ultimately emerges cannot be fully predetermined by the curatorial team. It develops through negotiation among professional knowledge, artistic context, and participant experience. This structure corresponds to Lacy's (1995) understanding of artistic practice as a platform for social relations and public exchange and demonstrates how public programming can extend from exhibition interpretation into cultural mediation and the organization of public issues.

The program also highlights an easily overlooked issue in participatory curating: visible activity and actual agency do not correspond perfectly. Listening cannot simply be defined as passive, and speaking publicly does not automatically constitute co-creation. A participant may critically reorganize their understanding while listening without expressing that process publicly; conversely, frequent speaking does not necessarily mean that a person's contribution genuinely enters the formation of shared meaning. For discourse-centered programs, the design of participatory mechanisms is therefore particularly important. Open-ended questions, direct participant-to-participant responses, time for disagreement, and methods for recording emergent themes can all make the audience's position within meaning production more explicit.

Compared with the first two projects, Pets' Impact therefore places greater demands on facilitation and discussion structure. Material works and personal narratives possess relatively visible forms, whereas discursive contributions can easily disappear when an event ends if they lack mechanisms for documentation, response, or recirculation. If the project is implemented, devices such as a question wall, discussion records, keyword mapping, or a post-event public text could allow viewpoints generated on site to re-enter the curatorial space. This would not only strengthen the visibility of discursive co-creation but also create analytical material for subsequent audience research.

5. Discussion

5.1 Participation Is Not a Single Category

Cross-case comparison shows that the differences among the three programs are not primarily a matter of "more" or "less" participation, but of the positions participants occupy within meaning production. Analysis of participatory public programs therefore needs to shift from activity formats to role relations: who can generate content, who can interpret it, who can respond to the contributions of others, and whether these contributions alter what becomes visible or discussable within the program. Participation is thus better understood as a set of

relationships organized by curatorial structures rather than as a single category.

This distinction matters because a formal increase in participatory activity does not necessarily mean that interpretive power has changed. If participant input were removed, would the program still communicate almost exactly the same content and interpretations? If the answer is yes, participation remains primarily a response to an existing structure. If participant-generated works, narratives, or opinions alter what others see, how they understand it, or what they discuss, the program moves closer to co-creation. This criterion places emphasis on whether participant contributions have "meaning consequences," rather than on whether an activity appears sufficiently interactive.

Accordingly, the article does not treat interaction, participation, and co-creation as a simple hierarchy of value, but as different configurations of agency. The crucial issue is not whether audiences act, but whether their actions enter chains of content production, interpretation, response, and circulation, and whether the institution preserves a meaningful place in which those contributions can have consequences.

5.2 Analytical Significance of the Three Modes of Co-Creation

The value of the three modes of co-creation lies not in relabeling projects but in making the question "what is being co-created?" more precise. Material, narrative, and discursive co-creation correspond respectively to different carriers of meaning: works and visible content; stories and interpretations; and public discussion and the negotiation of values. Because a single program may contain several of these carriers, the framework should be understood as a set of analytical dimensions rather than as mutually exclusive program categories.

The framework first separates co-creation from a narrow understanding centered on the production of material outcomes. Cultural institutions often favor outputs that can be photographed, displayed, and quantified, but visible products do not directly indicate how much interpretive freedom participants have acquired. Conversely, a discussion that produces no physical object may still possess a strong co-creative quality if participant viewpoints influence the direction of the issue and the understanding of others. Evaluation should therefore move from asking "how many outcomes were produced?" to asking "what can participants produce, change, and put into circulation?"

Cross-case comparison also highlights the importance of the circulation of meaning. Participant contributions acquire fuller public-cultural significance when they can be encountered, responded to, or reinterpreted by others. Family drawings depend on sharing and display to enter public space; youth photography gains social meaning through narrative and peer response; and adult discussion requires documentation, response, or follow-up texts if on-site viewpoints are not to disappear when the event ends. Co-creation concerns not only production but also how contributions become visible and acquire an afterlife within a temporary public.

The three modes therefore do not provide participatory curating with a hierarchical scale. Rather, they help curators distinguish the carriers of contribution, the ways contributions enter public space, and the interpretive consequences they may have for other participants.

5.3 Designing Participation for Different Publics

Comparison of the three projects also shows that audience segmentation acquires genuine curatorial significance only when it actually changes participatory mechanisms. Demographic information such as age, occupation, or interest does not automatically translate into effective public-program design. More important is how curators design corresponding entry pathways in relation to the media forms, expressive experience, and conditions of participation with which different publics are likely to be familiar.

The comparison further suggests that different publics do not require a linear increase in "openness," but different combinations of scaffolding and autonomy. The family program provides relatively explicit support through shared looking and making; the youth program leaves greater authorial space for image selection, narrative organization, and peer response; and the adult program requires a balance between the credibility of professional knowledge and the openness of discussion.

This also shows that participatory design does not mean minimizing structure. Different publics require different degrees of guidance and openness. For children and families, a completely open task may increase uncertainty; a clear sequence of looking, making, and sharing provides necessary scaffolding. For young participants, overly prescriptive creative tasks may restrict personal expression, so the program needs to leave greater space for authorship and peer interpretation. For adult discussion, a balance between professional knowledge and open exchange is required to prevent the event from reverting to one-way knowledge transmission.

Audience-responsive curating therefore does not mean giving every participant the same degree of freedom; it means adjusting structure to the modes of participation appropriate to particular publics. Emotional perception, creative expression, and critical reflection should likewise not be understood as a hierarchy of capacities among age groups. Adults can also enter an issue through emotion, children can make critical judgments, and young

people can participate in ethical debate. The three pathways describe only the principal invitation offered by each program, rather than essentializing audience capacities. Their function is to help curators ask how a particular public is most likely to enter a theme and how that initial form of participation might be guided toward more complex meaning-making.

5.4 From Audiences to Publics

Participatory public programming also produces an important conceptual shift: from understanding an "audience" as a pre-existing statistical category to understanding a "public" as a collectivity formed through circulation, response, and relationship. Warner's (2002) discussion emphasizes that publics are not defined solely in advance by demographic characteristics but emerge through the circulation of discourse and mutual response. For this study, public programming therefore involves not only recruiting target audiences but also creating conditions in which participants can see and respond to one another around a shared issue and form temporary cultural relationships.

This distinction is particularly significant in the case of The Pet Show. Companion animals are closely embedded in everyday life, and participants are likely to enter the programs with existing personal experiences of companionship, care, loss, responsibility, pet ownership, and family relationships. One function of public programming is to place these private experiences in relation to artworks, professional knowledge, and the experiences of other participants. Exhibition meaning is therefore no longer derived exclusively from the institution and artists, but is negotiated among curatorial knowledge, artistic practice, and public experience.

Within this structure, a public is not a pre-existing group waiting to be served by a cultural institution; it is temporarily organized through exchange, response, and shared interpretation. Families form small relational networks through collaborative making; young participants form peer interpretive communities through images and stories; and the adult program seeks to establish a discursive public around care, animal welfare, and ethical responsibility. Although these public relationships may be temporary, their significance lies in the institution's shift from merely communicating content to audiences toward creating conditions for cultural relationships among different subjects.

The movement from spectator to co-creator therefore does not mean that curatorial authority is fully relinquished. Cultural institutions remain responsible for framing issues, activity ethics, participant safety, the selection of collaborators, and basic standards of knowledge. A more appropriate model is distributed agency: curators design the conditions of participation; artists and professionals contribute contextual knowledge; participants generate new content and interpretations through works, narratives, and discussion; and institutions support the visibility, response, and circulation of those contributions. Co-creation emerges not when institutions withdraw, but when the roles of different actors in meaning production are reorganized.

5.5 Implications for Cultural-Institutional Practice

At the practical level, the analysis can be translated into a participatory curatorial logic composed of public identification, contribution design, circulation of meaning, and institutional responsibility. First, when identifying target publics, cultural institutions should move beyond demographic categories and consider what kinds of participatory relationships different publics can reasonably be invited to enter. Second, program design should specify what participants can actually contribute—works, stories, interpretations, questions, judgments, discussion, or combinations of these. Only when participant roles are made concrete can participation move from an abstract principle to an operational curatorial mechanism.

Third, programs need corresponding mechanisms for the circulation of meaning. If participant-generated content cannot be seen, responded to, or reinterpreted by others, co-creation can easily remain at an individual level. Temporary displays, peer response, collective discussion, textual documentation, and subsequent communication can all become important components of a co-creative structure. Fourth, participatory curating must clarify the responsibilities that institutions cannot fully relinquish. Where child safety, animal welfare, content ethics, professional information, or the representation of others is involved, institutions must continue to assume responsibility for structural design and ethical oversight.

The framework can also provide more targeted dimensions for evaluating public programs. Conventional evaluation often emphasizes attendance, satisfaction, and completion rates. If co-creation is a program goal, however, evaluation should also ask whether participants understood their roles, whether their contributions entered public space, whether participants responded to the content of others, and whether the program generated interpretations that were not entirely predetermined by the institution. Compared with measuring the scale of participation alone, these indicators more closely reflect participatory curating's concern with agency and meaning production.

Finally, the design of the three programs suggests that public programming should not be treated merely as an educational or promotional addition appended after an exhibition has been completed. When participation is

incorporated early in the curatorial process, public programs can become part of the exhibition's interpretive structure, providing another pathway for meaning production beyond artworks themselves. For cultural institutions seeking to position themselves as spaces of social dialogue and public reflection rather than merely display, the value of participatory public programming lies not only in expanding audience numbers but in rethinking how institutions organize relationships among different actors and how publics can genuinely enter processes of cultural meaning-making.

6. Limitations and Future Research

The principal limitation of this study is that it analyzes the structures of proposed programs rather than actual processes of participation. It therefore cannot determine whether participants would in practice take up the roles made available by the designs, nor can it observe how facilitation, group power differences, silent participation, self-selection, or unexpected on-site factors might alter the quality of co-creation. In other words, the study can analyze "what forms of agency are enabled by design," but cannot demonstrate the extent to which such agency would be realized in practice.

Future implementation could test the framework through mixed methods. Pre- and post-event questionnaires could examine changes in participants' understanding of exhibition issues; on-site observation could record who speaks, who makes decisions, who responds to whom, and whether different contributions genuinely enter public discussion; semi-structured interviews could explore how participants understand their own roles; and content analysis of drawings, photographs, captions, and discussion records could compare the meanings actually generated by the three modes of co-creation. If research involves minors, participant images, or records of on-site discussion, appropriate research ethics approval, parental or guardian consent, and informed-consent procedures should be completed before implementation.

A second limitation concerns age segmentation itself. Although age can be useful for program design, it may obscure differences in prior cultural participation, physical or cognitive characteristics, language, family structure, experience of pet ownership, and confidence in self-expression. Future research could therefore examine whether the proposed modes of co-creation are better predicted by life stage, prior experience, participatory motivation, or preferred modes of expression rather than age alone. The framework is intentionally designed with sufficient flexibility to support such revisions.

A third limitation concerns context. The three programs were designed for a photography-focused cultural institution in Shanghai and a pet-themed exhibition; their transferability cannot be assumed. Comparative research could examine whether material, narrative, and discursive co-creation operate in similar ways in history museums, science centers, artist-run spaces, or community cultural projects. Such comparison would help determine whether the framework describes broader patterns in participatory curating or is primarily useful as a design tool for a particular case.

7. Conclusion

Through a design-oriented case analysis of three public programs developed around Fotografiska Shanghai's The Pet Show, this article has examined how curatorial structures can create conditions for audiences to move from viewing toward co-creation. Rather than treating "participation" as a general indicator of activity, the analysis has focused on what participants are invited to produce, how their contributions enter public space, and whether those contributions can be responded to and circulated further.

The comparative analysis identifies three dominant modes of co-creation. The family workshop emphasizes material co-creation by translating exhibition viewing into collaborative drawing and public display. The youth photography exchange emphasizes narrative co-creation by positioning participants simultaneously as photographers, storytellers, viewers, and peer interpreters. The adult program emphasizes discursive co-creation by bringing professional knowledge, embodied encounter, and audience experience into ethical dialogue.

More broadly, participatory curating should not be equated with "more activities" or "greater informality." Meaningful participation depends on whether program structures create conditions in which participant contributions can have interpretive consequences. The curator's role is therefore not simply to withdraw, but to shift from being the sole producer of interpretation toward becoming a designer of participatory conditions. In practice, co-creation can acquire clearer public significance when participants can recognize their own contributions, encounter those of others, and see those contributions enter processes of viewing, interpretation, or discussion.

For cultural institutions, the framework offers a practical alternative to a "one form of participation fits all" approach. Public programs can be organized around three interconnected questions: Who is this public, and through what means can it most readily enter the issue? What are participants invited to contribute? What structures can make those contributions visible, responsive, and capable of influencing subsequent understanding and discussion? These questions cannot guarantee successful participation, but they can make the curatorial logic

of participation analyzable, comparable, and open to empirical testing after implementation. In this sense, the movement from audience to co-creator is not a predetermined effect but a question to be answered jointly through design and research.

References

Azoulay, A. (2008). *The civil contract of photography*. Zone Books.

Bennett, J. (2005). *Empathic vision: Affect, trauma, and contemporary art*. Stanford University Press.

Bishop, C. (2012). *Artificial hells: Participatory art and the politics of spectatorship*. Verso.

Black, G. (2005). *The engaging museum: Developing museums for visitor involvement*. Routledge.

Carpentier, N. (2011). *Media and participation: A site of ideological-democratic struggle*. Intellect.

Fotografiska. (n.d.). Manifesto. https://www.fotografiska.com/manifesto

Fotografiska Shanghai. (n.d.-a). The Pet Show #20: Sophie Gamand. https://shanghai.fotografiska.com/en/the-pet-show-20-sophie-gamand

Fotografiska Shanghai. (n.d.-b). The Pet Show #22: Wang Hanlin. https://shanghai.fotografiska.com/en/the-pet-show-22-wang-hanlin

Hooper-Greenhill, E. (2000). *Museums and the interpretation of visual culture*. Routledge.

Kester, G. H. (2004). *Conversation pieces: Community and communication in modern art*. University of California Press.

Lacy, S. (Ed.). (1995). *Mapping the terrain: New genre public art*. Bay Press.

Lynch, B. T. (2011). Whose cake is it anyway? A collaborative investigation into engagement and participation in 12 museums and galleries in the UK. Paul Hamlyn Foundation.

Simon, N. (2010). The participatory museum. Museum 2.0.

Warner, M. (2002). Publics and counterpublics. *Public Culture, 14*(1), 49–90.

Copyrights