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.

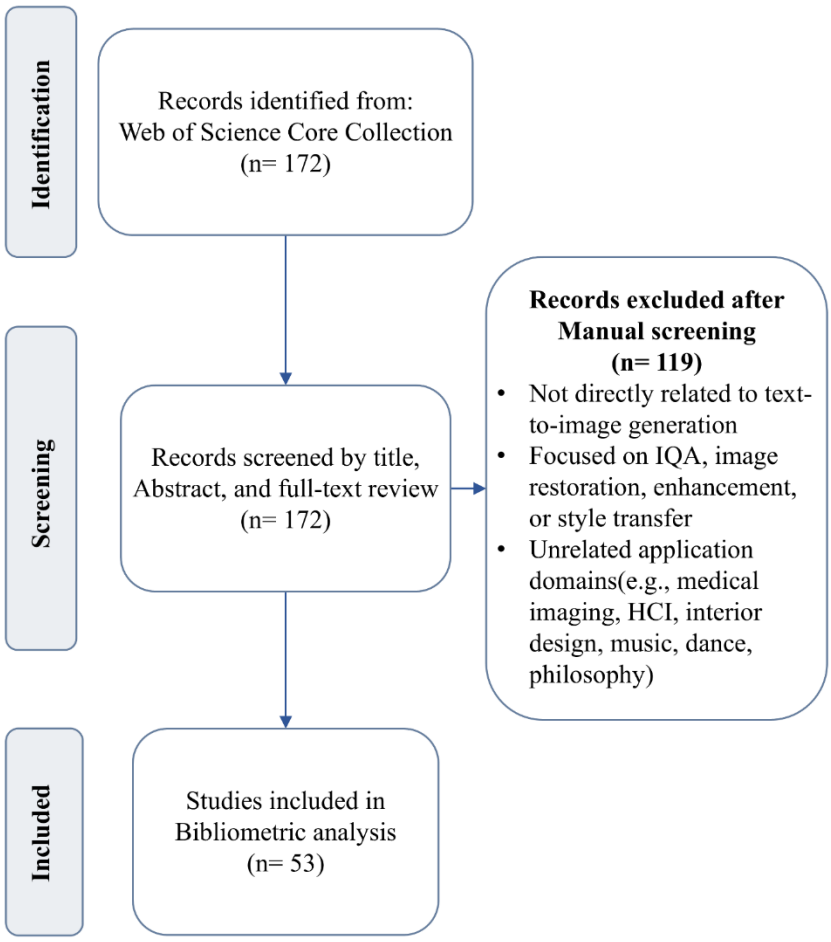

Figure 1. PRISMA 2020-based flow diagram of study selection

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.

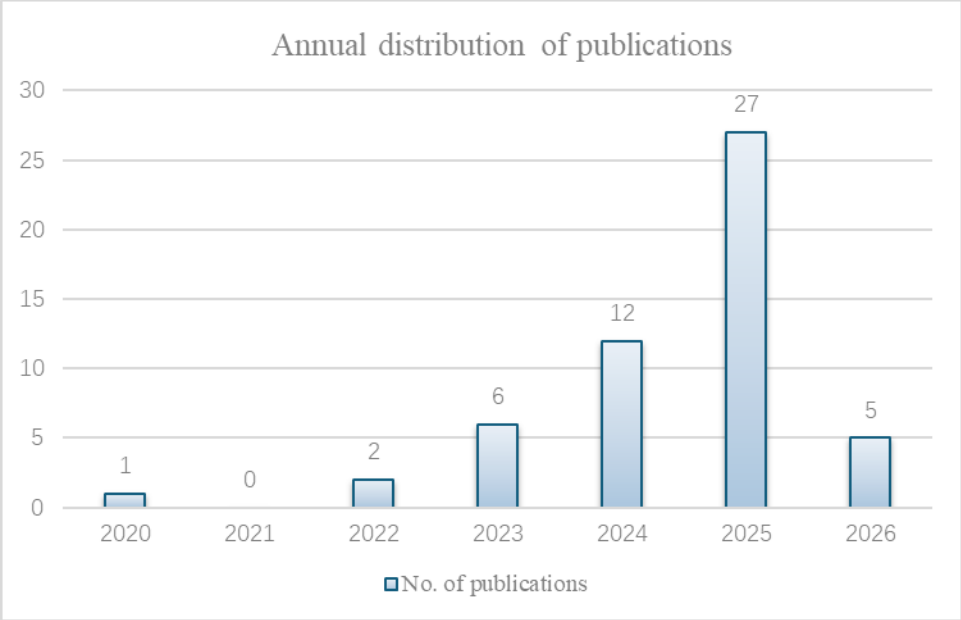

Figure 2. Annual distribution of publications

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.

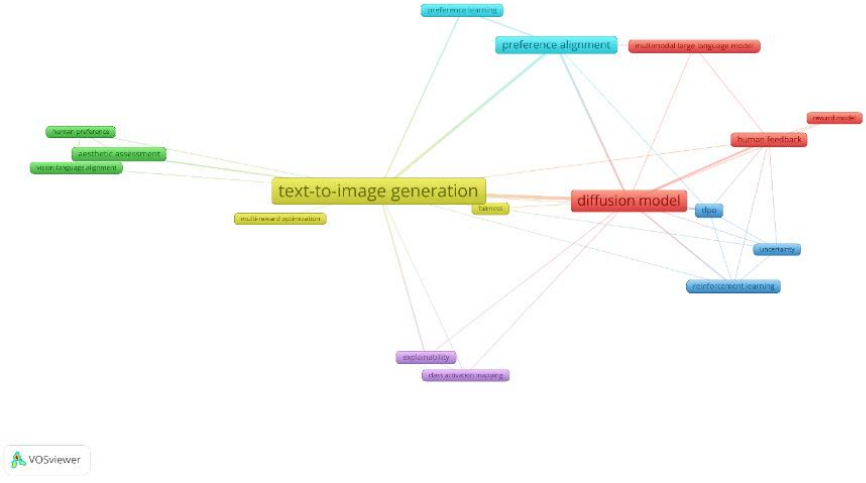

Figure 3. Keyword co-occurrence network of the T2I computational aesthetics corpus

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.

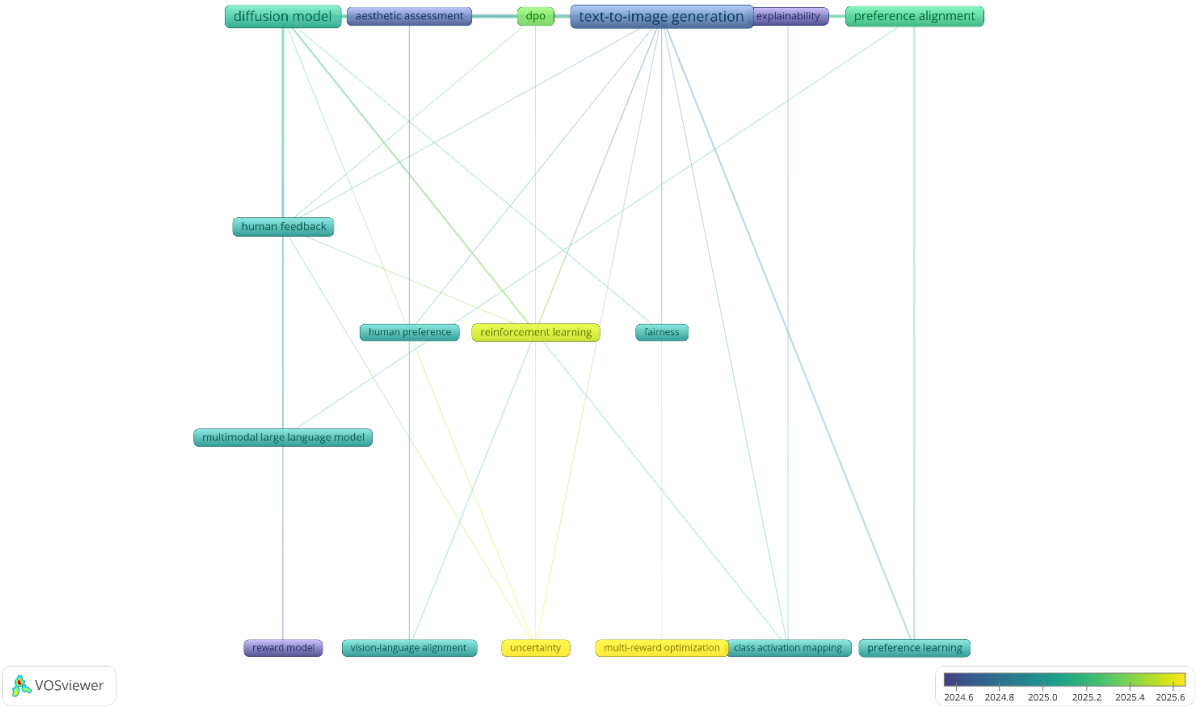

Figure 4. Temporal overlay map of keywords in the T2I computational aesthetics corpus

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.

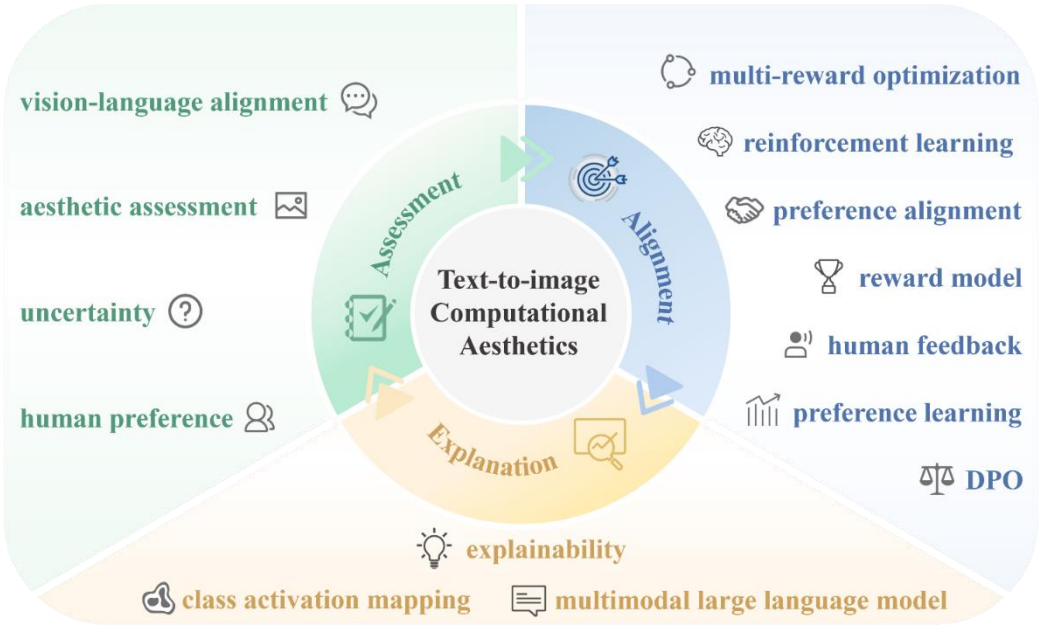

Figure 5. Keyword-informed conceptual framework of T2I computational aesthetics

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.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).