

# Voice-Over as a Substitute for Visual Explanation in Short Videos

M. D. Reynolds<sup>1</sup>

<sup>1</sup> University of Guelph, Canada

Correspondence: M. D. Reynolds, University of Guelph, Canada.

doi:10.63593/JLCS.2025.12.04

## Abstract

Short-form video has become a dominant mode of audiovisual communication on contemporary digital platforms, where limited duration, rapid editing, and fragmented viewing practices increasingly undermine the explanatory capacity of visual continuity. This paper examines the growing role of voice-over narration as a substitute for visual explanation in short videos and argues that voice-over should be understood not as a stylistic or technical supplement but as a core explanatory mechanism. Drawing on audiovisual theory, narration studies, and platform media research, the analysis shows that visual explanation traditionally depends on temporal development, spatial coherence, and sustained attention, conditions that are structurally weakened in short video formats. In response, voice-over assumes primary explanatory authority by guiding interpretation, condensing processes, and stabilizing meaning across fragmented visuals, offering a level of clarity and abstraction that images alone struggle to achieve under accelerated and distracted consumption. The paper situates this shift within broader media conditions such as everyday multitasking, the perceptual stability of sound, platform norms favoring rapid comprehension, and cultural preferences for explicit guidance, and it discusses the narrative and cultural implications of this transformation, including a movement from showing to telling, reduced interpretive openness, and the normalization of guided meaning. By reframing voice-over as a substitute for visual explanation, the study challenges assumptions of visual dominance in audiovisual media and highlights a rebalancing of sound and image in platform-based communication.

**Keywords:** voice-over narration, short-form video, visual explanation, audiovisual theory, platform media, guided meaning

## 1. Introduction

### 1.1 Background

Short-form video platforms have become a central infrastructure of contemporary media consumption and everyday communication. They are no longer marginal or supplementary to traditional media forms but operate as a

primary interface through which information, knowledge, and cultural meanings circulate. The defining characteristics of short videos include extreme temporal compression, algorithmically driven distribution, and an emphasis on immediate perceptual impact. These conditions shape not only what kind of content is produced but also how meaning is constructed and

received.

Within this environment, visual material is subject to intense pressure. Images are required to attract attention instantly, maintain engagement within seconds, and remain legible on small mobile screens. Editing practices privilege speed and rhythm over continuity, while narrative structures are often reduced to highlights, summaries, or isolated moments. As a result, visuals are frequently fragmented and decontextualized. They present effects without causes, outcomes without processes, and moments without extended temporal development. The traditional expectation that images can explain through duration, spatial coherence, and sequential logic becomes increasingly difficult to sustain.

At the same time, short videos circulate within conditions of everyday distraction. Viewing often occurs alongside other activities, with intermittent visual attention and fluctuating levels of cognitive engagement. Under such circumstances, visual meaning alone becomes unstable. Images may be seen but not fully processed, or they may be skipped altogether while audio continues to play. These viewing practices further undermine the explanatory capacity of visuals and create demand for alternative channels of meaning transmission.

It is within this media ecology that voice-over narration has become pervasive. Across diverse genres, spoken language is used to explain, summarize, interpret, and frame visual content. Voice-over no longer appears as an optional stylistic layer added to otherwise self-sufficient images. Instead, it increasingly functions as a structural component that stabilizes meaning in an environment where visual explanation is weakened by fragmentation, speed, and distraction.

### 1.2 Research Focus

This paper approaches voice-over not as a technical feature or a narrative embellishment but as a functional response to the changing conditions of audiovisual meaning-making. The focus lies on why voice-over increasingly operates as a substitute for visual explanation rather than a supplement to it. The analysis centers on the relationship between fragmented visuals and spoken narration, asking how voice-over compensates for what images can no longer reliably convey on their own.

Rather than examining production techniques or

conducting empirical audience studies, the paper adopts a theoretical and conceptual perspective. It seeks to understand how explanatory labor is redistributed between sound and image in short video formats. Attention is directed toward the ways spoken language guides interpretation, condenses information, and resolves ambiguity in contexts where visuals are insufficiently continuous or complete.

The core question is not whether voice-over improves clarity but why clarity itself has become increasingly dependent on speech. By situating voice-over within broader platform conditions such as acceleration, attention scarcity, and algorithmic optimization, the paper frames narration as an adaptive mechanism shaped by structural constraints rather than individual creative choice.

### 1.3 Research Significance

Conceptualizing voice-over as an explanatory mechanism has implications for how short videos are understood within media theory. Audiovisual studies have often privileged the image as the dominant carrier of meaning, with sound positioned as supportive or secondary. The prominence of voice-over in short videos challenges this hierarchy and calls for a reconsideration of how meaning is organized across sensory channels in contemporary media.

This perspective contributes to discussions on narration, perception, and platform culture by highlighting a shift in explanatory authority from visual demonstration to verbal articulation. It suggests that short video platforms foster forms of communication in which meaning is increasingly guided, condensed, and stabilized through speech. Such a shift has consequences for audience interpretation, narrative openness, and visual literacy.

By focusing on voice-over as a substitute rather than an addition, this study provides a framework for understanding broader transformations in audiovisual communication under platform conditions. It opens space for further inquiry into how sound, voice, and language reassert their centrality in media environments often assumed to be dominated by images.

## 2. Visual Explanation and Its Limits in Short Videos

### 2.1 Visual Explanation in Audiovisual Media

Visual explanation has long been regarded as a foundational capacity of audiovisual media. In cinema, television, and documentary traditions, images are expected to carry explanatory weight through their organization in time and space. Visual meaning emerges from the ability of images to show relations rather than simply present isolated objects or moments. Processes become intelligible when actions unfold gradually, when changes can be observed across duration, and when spatial continuity allows viewers to orient themselves within a scene. Explanation is achieved not by stating facts directly but by arranging visual elements so that understanding arises through observation.

This mode of explanation presupposes a certain stability in viewing conditions. Images are given time to develop, and viewers are assumed to maintain sustained attention. Continuity editing, establishing shots, and sequential framing all contribute to a visual logic in which causes precede effects and actions lead to consequences. Even when audiovisual works rely on montage or symbolic imagery, explanation still depends on the accumulation of visual cues that can be connected through interpretive effort.

Visual explanation also relies on the contextual richness of images. Background details, environmental cues, and bodily gestures provide information that exceeds explicit narration. A scene can communicate social relations, emotional states, or power dynamics without verbal clarification because viewers are able to scan the image, notice details, and integrate them into a coherent understanding. In this sense, visual explanation depends on openness and density rather than speed. The image offers more information than is immediately consumed, allowing meaning to emerge gradually.

Another key feature of visual explanation is its reliance on inferential participation. Viewers are not simply recipients of information but active interpreters who construct meaning by linking visual elements together. Explanation is distributed across shots and sequences rather than concentrated in a single moment. This model assumes that viewers are willing and able to invest cognitive effort in interpretation. Visual explanation therefore aligns with media forms that value immersion, narrative depth, and interpretive engagement.

## 2.2 Structural Constraints of Short-Form Video

Short-form video platforms introduce a radically different set of structural conditions that challenge the assumptions underlying traditional visual explanation. Duration is the most obvious constraint. When a video lasts only a few seconds or minutes, there is little room for gradual development. Processes must be compressed or omitted, and visual sequences are often reduced to highlights rather than complete representations. The image no longer unfolds but appears in fragments that are immediately replaced by the next visual stimulus.

Editing practices in short videos prioritize rhythm and immediacy. Rapid cuts, jump edits, and abrupt transitions are common, not as stylistic experimentation but as normative conventions. Spatial coherence is frequently disrupted as scenes shift locations without establishing context. Temporal continuity is often abandoned, with past, present, and future moments collapsed into a single sequence. Under these conditions, images struggle to establish causal relations. Actions may be shown without sufficient buildup, and outcomes may appear without clear explanation of how they were achieved.

The visual economy of short videos is also shaped by platform competition. Countless videos vie for attention within algorithmically curated feeds, encouraging creators to maximize visual impact in the shortest possible time. Images are designed to capture attention instantly rather than sustain understanding. Visual shock, novelty, and emotional intensity are rewarded, while slow exposition is penalized by the risk of being skipped. This incentive structure discourages visual explanation that requires patience or sustained observation.

Screen size and viewing environment further complicate visual explanation. Short videos are primarily consumed on mobile devices, often in vertical formats that limit visual scope. Fine details may be difficult to perceive, and complex spatial arrangements may lose clarity. Viewing conditions are rarely controlled, with ambient distractions competing for attention. Visual explanation that depends on subtle cues or careful observation becomes fragile under such circumstances.

Short-form video narratives are also shaped by

repetition and modularity. Many videos follow similar templates, formats, or trends, leading to visual standardization. When images become predictable, they lose explanatory specificity. A gesture, setting, or visual trope may signal a genre or emotion without providing concrete information about a particular situation. Visual meaning becomes shorthand rather than explanation, relying on prior familiarity rather than visual demonstration.

### 2.3 *The Weakening of Visual Self-Sufficiency*

These structural constraints contribute to a weakening of visual self-sufficiency in short videos. Visual self-sufficiency refers to the capacity of images to explain themselves through their internal organization and contextual richness. In short video formats, this capacity is undermined by fragmentation, speed, and decontextualization. Images are no longer expected to carry meaning independently but function as partial signals that require external support.

One consequence of this shift is the separation of attention from comprehension. Viewers may register visual stimuli without fully understanding them. An image can be seen, recognized, or emotionally felt without being cognitively processed in depth. Visual attention becomes fleeting, while understanding requires additional cues that images alone cannot reliably provide. The gap between perception and comprehension widens as images circulate faster than they can be interpreted.

Another consequence is the reduction of visual causality. Short videos often present results rather than processes. A transformation, achievement, or outcome is shown without visual evidence of how it occurred. The explanatory chain is broken, leaving images as endpoints rather than narratives. In such cases, visuals attract curiosity but fail to satisfy it. They prompt questions that cannot be answered visually within the available time and structure.

Visual ambiguity also takes on a different role. In longer audiovisual forms, ambiguity can invite interpretation and deepen engagement. In short videos, ambiguity is often experienced as confusion rather than openness. When viewers lack time or context to resolve ambiguity, unclear images risk disengagement. The tolerance for interpretive uncertainty decreases, and the demand for immediate clarity increases. Visual explanation struggles under these

conditions because it traditionally relies on interpretive openness and gradual meaning formation.

The weakening of visual self-sufficiency also affects the authority of images. When visuals cannot fully explain, they lose their status as primary sources of meaning. They become dependent on external framing to anchor interpretation. This dependence does not necessarily diminish the aesthetic appeal of images, but it changes their function. Images shift from being explanatory structures to being illustrative elements that support meaning articulated elsewhere.

This transformation is closely linked to changes in audience expectations. Viewers accustomed to narrated short videos may no longer approach images with the assumption that they should be interpreted independently. Instead, they may wait for verbal cues to clarify meaning. Visual literacy practices adapt to platform norms, and the skill of extracting explanation from images alone becomes less central. Visual self-sufficiency declines not only because of structural limitations but also because of changing habits of interpretation.

The limits of visual explanation in short videos therefore cannot be attributed solely to technical constraints. They reflect a broader reconfiguration of audiovisual communication. Images remain powerful in capturing attention and evoking emotion, but their capacity to explain is increasingly constrained by the conditions under which they are produced, distributed, and consumed. Visual explanation, once a defining strength of audiovisual media, becomes unreliable in environments characterized by acceleration, fragmentation, and partial attention.

This does not imply the disappearance of visual meaning. Instead, it suggests a redistribution of explanatory labor. Visuals continue to play an important role, but they no longer operate as autonomous explanatory systems. Their meaning is increasingly stabilized through other channels, particularly spoken language. Understanding the limits of visual explanation in short videos is therefore a necessary step toward understanding why voice-over emerges as a substitute rather than a supplement in contemporary platform media.

## 3. **Voice-Over as an Explanatory Mechanism**

### 3.1 *Voice-Over as Semantic Guidance*



Voice-over operates as a direct mechanism for organizing meaning in short videos by providing explicit semantic guidance. Spoken language has the capacity to name, classify, evaluate, and explain in ways that images alone often cannot under conditions of fragmentation and speed. Through voice-over, creators are able to establish what a scene is about, which elements are relevant, and how the viewer should interpret what is shown. This guidance does not emerge gradually but is delivered immediately, often within the first seconds of a video, shaping interpretation before the visuals have had time to develop their own logic.

Semantic guidance through voice-over functions by reducing uncertainty. Images in short videos frequently present incomplete information. A scene may show an action without context, a result without process, or an object without explanation. Voice-over fills these gaps by supplying missing links. It can state causal relations that are visually absent, define abstract concepts that cannot be shown directly, or situate an image within a broader narrative or argument. In doing so, voice-over transforms visual fragments into legible units of meaning.

This guiding function also operates at the level of attention. Voice-over directs viewers toward specific aspects of the image, telling them what to notice and what to ignore. In a visually dense or rapidly changing scene, not all elements can be processed equally. Spoken narration prioritizes certain details by naming them or assigning significance to them. The viewer's gaze is guided not by visual exploration but by verbal instruction. Interpretation becomes aligned with the logic of speech rather than the openness of the image.

Voice-over guidance also structures emotional interpretation. Images can evoke multiple emotional responses depending on context and personal experience. Voice-over narrows this range by framing how a scene should be felt. A visual moment that might otherwise appear neutral or ambiguous can be defined as inspiring, shocking, tragic, or humorous through narration. Emotional meaning becomes less dependent on visual cues and more dependent on verbal framing.

In this sense, voice-over does not merely add information to visuals. It actively organizes perception. It establishes a hierarchy of meaning in which spoken language occupies a directive

role, shaping how images are seen, understood, and evaluated. Semantic guidance through voice-over reflects a shift in explanatory authority away from visual inference toward verbal instruction.

### 3.2 *Substitution Rather Than Supplementation*

In traditional audiovisual theory, voice-over is often described as supplementary. It is seen as an additional layer that supports or enhances visual meaning without replacing it. In short video practices, this relationship is increasingly reversed. Voice-over becomes the primary explanatory channel, while visuals function as secondary elements that illustrate or reinforce what is already explained verbally.

This substitution occurs because visuals in short videos often lack the conditions required for autonomous explanation. Fragmented shots, compressed timelines, and minimal context limit the ability of images to construct meaning independently. Rather than attempting to restore visual continuity, creators rely on voice-over to carry the explanatory burden. Spoken narration provides the structure that images can no longer sustain.

In many short videos, the narrative logic unfolds entirely through speech. The voice-over introduces the topic, explains its significance, and delivers conclusions. Visuals appear as loosely connected scenes that accompany the narration but do not determine it. If the audio were removed, the video would often become difficult or impossible to understand. If the visuals were removed, the core message might still remain intelligible. This asymmetry indicates a clear substitution rather than mutual dependence.

The substitution of voice-over for visual explanation also reshapes the function of images. Images become illustrative rather than demonstrative. They serve to maintain viewer interest, provide sensory stimulation, or evoke emotion, while the task of explanation is handled by language. Visuals no longer need to show how something works or why it happens. They only need to align loosely with what is being said. This reduces the demand placed on visual coherence and allows for greater flexibility in image selection.

This shift also affects narrative hierarchy. In classical audiovisual storytelling, images often lead and sound follows. In short videos with dominant voice-over, sound leads and images

follow. Visual sequences are edited to match the rhythm and content of speech rather than the other way around. The narrative spine is verbal, and the image track adapts to it. This inversion reflects a broader transformation in how audiovisual meaning is assembled under platform constraints.

Substitution rather than supplementation also alters the viewer's interpretive role. When explanation is carried by voice-over, viewers are less required to infer meaning from visual relationships. Interpretation becomes a process of listening and accepting rather than observing and connecting. Visual engagement remains present, but it is guided and constrained by verbal explanation. The image loses some of its autonomy as a site of meaning production.

### *3.3 Efficiency of Verbal Explanation*

The efficiency of verbal explanation is a key reason for the prominence of voice-over in short videos. Spoken language is capable of compressing complex information into concise statements. Causal chains, abstract ideas, and evaluative judgments can be expressed in seconds. Visual explanation of the same content would require extended sequences, multiple shots, and careful pacing, which short video formats rarely allow.

Language excels at summarization. A voice-over can condense an entire process into a brief verbal description, bypassing the need to show intermediate steps. This is particularly valuable in educational, informational, or commentary-based short videos, where the goal is to transmit knowledge quickly. Visual demonstration of a process may be incomplete or misleading if shown too briefly. Verbal explanation avoids this risk by stating conclusions directly.

Verbal explanation is also efficient in handling abstraction. Many concepts addressed in short videos are not easily visualized. Ideas related to motivation, social trends, ethical judgments, or personal reflection often lack concrete visual referents. Voice-over allows creators to address such topics without relying on symbolic or metaphorical imagery, which may be difficult to interpret quickly. Speech provides clarity where visuals might introduce confusion.

The temporal flexibility of voice-over contributes to its efficiency. Speech can overlap with visuals, allowing explanation and imagery to coexist within the same time frame. Visual

explanation often requires sequential time. A process must be shown step by step. Voice-over can explain while something else is being shown, maximizing the use of limited duration. This parallelization of information channels makes voice-over particularly suited to short formats.

Efficiency also relates to cognitive load. Visual explanation often requires viewers to actively process relationships between images. This demands attention and interpretive effort. Voice-over reduces this burden by delivering meaning directly. Viewers can understand the message without reconstructing it themselves. In environments characterized by distraction and rapid consumption, this reduction in cognitive effort is advantageous.

The efficiency of verbal explanation aligns with platform incentives. Algorithms favor content that is quickly understood and easily consumed. Voice-over supports these goals by ensuring that meaning is accessible even during brief viewing moments. Speech can convey the main point early, increasing the likelihood that viewers will continue watching or engage with the content.

Efficiency, however, is not a neutral quality. While it enables communication under constraints, it also shapes the form of meaning that is communicated. Efficient verbal explanation tends to favor clarity over complexity and conclusion over exploration. It supports content that delivers answers rather than raises questions. The dominance of voice-over as an explanatory mechanism reflects a broader orientation toward speed and legibility in platform media.

Taken together, semantic guidance, substitution of visual explanation, and the efficiency of verbal language explain why voice-over has become central to meaning-making in short videos. Voice-over is not merely a response to technical limitations but a structural adaptation to the conditions of contemporary media consumption. It reorganizes the relationship between sound and image, redefining how explanation is achieved in compressed audiovisual forms.

## **4. Media Conditions Favoring Voice-Over Explanation**

### *4.1 Partial Attention in Everyday Media Use*

Short-form video consumption is deeply embedded in everyday routines rather than

separated as a dedicated viewing activity. Unlike cinema or television, which traditionally require a fixed viewing position and a relatively stable attentional state, short videos are consumed in transit, during breaks, and alongside other tasks. Viewing often takes place while commuting, waiting, eating, or switching between applications. In these situations, attention is divided and unstable. The screen may be glanced at briefly, then ignored, then returned to, creating a fragmented pattern of visual engagement.

Under such conditions, visual explanation becomes unreliable. Visual meaning presupposes sustained attention, even when images are simple or familiar. Understanding visual relations requires looking, comparing, and integrating what is seen across time. Partial attention interrupts this process. When viewers miss key visual moments or fail to track visual continuity, images lose their explanatory power. They remain perceptible as stimuli but fail to cohere into intelligible sequences.

Audio operates differently in distracted contexts. Sound can be received without direct orientation toward the screen. A viewer may hear narration while looking elsewhere, performing another task, or momentarily disengaging from the video interface. Voice-over remains accessible even when visual attention drops to a minimum. This asymmetry between visual and auditory attention makes sound a more reliable channel for meaning transmission in everyday media use.

Voice-over aligns with this attentional reality by ensuring that the core message does not depend on uninterrupted viewing. Spoken narration allows meaning to persist across moments of visual disengagement. A viewer who looks away from the screen may still follow the argument, explanation, or story through audio alone. When attention returns to the screen, visuals can be reintegrated without having been solely responsible for explanation.

Partial attention also shapes cognitive expectations. Viewers accustomed to distracted viewing may no longer expect to derive meaning primarily from images. Instead, they rely on narration to anchor understanding. Visuals become supplementary cues that enhance or illustrate what is already understood verbally. The explanatory burden shifts toward voice because it accommodates the realities of

divided attention more effectively than visual sequences.

This pattern does not indicate a decline in visual interest or appreciation. Images remain central to attracting attention and maintaining engagement. Their role, however, changes from explaining to accompanying. In a media environment structured around partial attention, voice-over becomes the channel through which coherence is maintained despite interruptions, distractions, and fluctuating focus.

#### *4.2 Sound as a Stable Perceptual Channel*

Sound possesses perceptual characteristics that make it particularly suited to explanatory functions in short video environments. Auditory perception unfolds over time in a continuous manner. Speech remains intelligible as long as it is heard, regardless of whether the listener maintains visual contact with the source. Images, by contrast, depend on spatial fixation and visual orientation. When the screen is not actively watched, visual information ceases to be available.

The stability of sound lies in its temporal persistence. Voice-over flows continuously, creating a stable thread of meaning that runs through rapidly changing images. Even when visuals shift abruptly, narration provides continuity. This continuity allows viewers to maintain a sense of progression and coherence despite visual fragmentation. Spoken language connects moments that might otherwise appear disjointed.

Sound is also less susceptible to perceptual overload. Visual environments in short video platforms are highly saturated. Bright colors, rapid motion, text overlays, and transitions compete for attention. This density can overwhelm visual processing, making it difficult to extract explanatory information. Voice-over cuts through visual noise by occupying a different sensory channel. It simplifies the interpretive task by delivering meaning directly, reducing the need to decode complex visual arrangements.

Auditory explanation also benefits from familiarity. Human speech is one of the most deeply learned and practiced forms of communication. Listeners are skilled at processing spoken language quickly, even under suboptimal conditions. Accents, background noise, and variable audio quality can be

accommodated with relative ease. This robustness enhances the reliability of voice-over as an explanatory medium in uncontrolled viewing environments.

The stability of sound also extends to emotional and narrative continuity. A consistent voice can create a sense of presence and authority that anchors the viewer's experience. Even when visuals change style, location, or pace, the voice remains constant. This constancy supports narrative cohesion and reinforces the explanatory role of narration. The voice becomes a guide through visual variation rather than a supplement to it.

In short video contexts, where images are often selected for impact rather than coherence, sound assumes the task of stabilizing meaning. Voice-over does not merely accompany visuals but provides a continuous interpretive framework that compensates for visual volatility. This perceptual stability makes sound an ideal channel for explanation under conditions of rapid visual change.

#### 4.3 Platform Norms and Audience Expectations

Platform environments play a crucial role in normalizing voice-over as an explanatory strategy. Short video platforms are not neutral spaces but structured ecosystems shaped by algorithms, content trends, and user behavior. Over time, certain formats and conventions become dominant because they perform well within these systems. Narrated videos are among the most prominent of these conventions.

As voice-over becomes widespread, audiences develop expectations for verbal explanation. Viewers learn to anticipate narration as a guide to meaning. When a video lacks voice-over, it may be perceived as incomplete, confusing, or less accessible. This expectation is not necessarily conscious but emerges through repeated exposure to narrated content. Platform cultures gradually define what a legible video looks and sounds like.

Creators respond to these expectations by incorporating voice-over as a default element. Even when visuals could potentially explain on their own, narration is added to ensure clarity and alignment with audience habits. Over time, this reinforces the association between explanation and speech. Voice-over becomes institutionalized as a standard feature rather than a creative choice.

Platform metrics further reinforce this trend. Videos that are quickly understood and easily consumed tend to perform better in terms of engagement, retention, and sharing. Voice-over supports these metrics by reducing interpretive effort and accelerating comprehension. Algorithms indirectly favor narrated content by rewarding clarity, which encourages creators to rely on verbal explanation.

Audience expectations also intersect with accessibility. Voice-over can make content more inclusive for viewers with limited visual attention, small screens, or situational constraints. While captions address accessibility for hearing-impaired viewers, voice-over addresses the inverse situation of limited visual access. Platforms that prioritize broad reach implicitly favor content that can be understood through multiple sensory channels, with sound playing a central role.

As narration becomes normalized, visual explanation is no longer the default standard against which videos are judged. Instead, explanation through voice is taken for granted, and images are evaluated based on their ability to complement narration rather than replace it. Platform norms thus reshape the criteria of successful communication, privileging guided meaning over visual inference.

#### 4.4 Algorithmic Time Pressure and Accelerated Comprehension

Short video platforms operate under conditions of extreme temporal competition. Users scroll rapidly through feeds, often spending only seconds on each piece of content. Videos must establish relevance and meaning almost immediately to prevent being skipped. This creates intense time pressure at the level of perception and interpretation.

Visual explanation struggles under such pressure. Visual meaning often requires time to unfold. Establishing context, showing relationships, and allowing viewers to observe changes across duration cannot be easily compressed into the first few seconds. Voice-over, by contrast, can state the main point immediately. A single sentence can define the topic, summarize the outcome, or frame the narrative direction.

Algorithmic systems reward this immediacy. Content that communicates its purpose quickly is more likely to retain viewers and be promoted further. Voice-over allows creators to front-load



explanation, ensuring that viewers understand what the video is about before deciding whether to continue watching. Visual explanation alone cannot always achieve this level of instant clarity.

Accelerated comprehension also affects how content is structured. Videos are increasingly designed around clear verbal hooks, summaries, and conclusions. Voice-over provides these elements efficiently. The image track adapts to this structure by offering illustrative scenes rather than explanatory sequences. The explanatory core becomes verbal, optimized for speed and legibility.

This acceleration reshapes the experience of meaning itself. Understanding is no longer something that unfolds gradually through observation but something that is delivered quickly through narration. The role of the viewer shifts from interpreter to receiver. Voice-over fits seamlessly into this accelerated mode of comprehension, while visual explanation appears slow and uncertain by comparison.

#### 4.5 Multitasking and the Audio Background Function

Short videos often function as background media rather than foreground experiences. Users may play videos while doing other activities, treating audio as a continuous stream that accompanies daily routines. In such cases, visuals may be only intermittently consulted or ignored entirely.

Voice-over enables this mode of consumption by transforming short videos into quasi-audio content. Explanation delivered through speech allows the video to remain meaningful even when not actively watched. This flexibility expands the contexts in which short videos can be consumed and increases their integration into everyday life.

The background function of audio also influences content design. Creators may prioritize narration that remains intelligible without visuals. Explanations are phrased clearly, references to specific visual details are minimized, and meaning is made explicit through speech. Visuals become optional enhancements rather than essential components of understanding.

This shift has implications for how media presence is experienced. Voice-over creates a

sense of companionship or guidance that persists even when attention is divided. The voice accompanies the viewer through other activities, reinforcing its explanatory authority. Visual explanation, which requires focused attention, is less compatible with this mode of engagement.

#### 4.6 Cultural Preferences for Explicitness and Guidance

Media conditions also reflect broader cultural shifts toward explicit communication and guided interpretation. In fast-paced information environments, ambiguity is often experienced as inefficiency rather than openness. Viewers seek content that delivers clear messages, actionable insights, or straightforward narratives.

Voice-over satisfies this preference by making meaning explicit. It reduces the need for inference and minimizes interpretive uncertainty. Visual explanation, which often relies on implication and gradual understanding, may appear vague or demanding under such expectations.

This cultural orientation toward explicitness reinforces the role of narration as an explanatory tool. Voice-over aligns with a desire for clarity and immediacy, while visual explanation may be perceived as indirect or incomplete. Media conditions thus favor modes of communication that prioritize telling over showing.

Taken together, these media conditions explain why voice-over has become central to explanation in short videos. Partial attention, perceptual stability of sound, platform norms, algorithmic time pressure, multitasking practices, and cultural preferences converge to privilege spoken narration. Voice-over is not simply added to short videos but emerges as a structural solution to the conditions under which they are produced and consumed.

### 5. Narrative and Cultural Implications

#### 5.1 From Showing to Telling

The increasing reliance on voice-over in short videos signals a profound narrative shift from showing to telling. In classical visual storytelling, meaning is expected to emerge through observation. Images unfold in time, allowing viewers to grasp relationships, motivations, and consequences by watching actions take place. The ideal of visual narration emphasizes demonstration over explanation, trusting the audience to infer meaning from

what is seen rather than what is said. Short video practices increasingly depart from this model.

Voice-over reorients narrative structure toward verbal articulation. Instead of allowing images to build meaning through sequence and duration, narration often states meaning directly. What happened, why it happened, and what it signifies are explained explicitly through speech. Visuals become evidence or illustration rather than the primary site of narrative development. This transformation alters the fundamental logic of storytelling. Narrative progression is no longer anchored in visual causality but in verbal sequencing.

This shift aligns short video storytelling with oral narrative traditions. Oral narration privileges immediacy, clarity, and direct address. The storyteller guides listeners through events, highlights relevant points, and interprets significance on their behalf. Meaning is transmitted through voice rather than discovered through observation. Short videos increasingly adopt this mode, positioning the narrator as a guiding presence who leads the audience through a compressed narrative experience.

The move from showing to telling also reflects changing assumptions about audience patience and interpretive labor. Visual storytelling requires time and attention. It assumes that viewers are willing to observe, compare, and infer. Telling assumes that viewers prefer direct explanation and quick comprehension. Voice-over satisfies this preference by delivering narrative meaning efficiently. In doing so, it redefines what counts as effective storytelling under platform conditions.

This transformation challenges long-standing ideals within media theory that value visual subtlety, ambiguity, and experiential immersion. When narration takes precedence, images lose their narrative autonomy. They no longer need to demonstrate processes or embody temporal development. Their role shifts toward supporting verbal claims. The narrative authority moves from the image track to the voice track, reshaping the hierarchy between sound and image.

The implications extend beyond individual videos. As telling becomes normalized, audiences adapt their expectations. Viewers may come to see visual storytelling without narration

as incomplete or unclear. The cultural standard of what a coherent story looks like changes. Storytelling becomes something that is explained rather than shown, spoken rather than visualized.

### *5.2 Reduced Interpretive Openness*

One consequence of narration-centered storytelling is a reduction in interpretive openness. Visual explanation traditionally allows for multiple readings. Images can be ambiguous, inviting viewers to project their own meanings, emotions, or experiences onto what they see. This openness is not a flaw but a defining strength of visual media. It enables richness, complexity, and personal engagement.

Voice-over narrows this openness by defining meaning in advance. When narration explains what an image represents or how it should be understood, alternative interpretations are discouraged. The viewer is guided toward a preferred reading, leaving less space for personal inference. Ambiguity is resolved quickly through verbal clarification. Meaning becomes fixed rather than negotiated.

This narrowing of interpretation affects the role of the audience. Instead of actively constructing meaning from visual cues, viewers are positioned as recipients of explanation. The interpretive work is performed by the narrator, not the viewer. Visual engagement becomes confirmatory rather than exploratory. Images are used to illustrate what has already been said rather than to provoke reflection or questioning.

Reduced interpretive openness also affects emotional experience. Images often evoke complex or mixed emotions that resist easy categorization. Voice-over tends to label emotional meaning, framing a scene as inspiring, tragic, ironic, or humorous. Emotional interpretation becomes guided by language. The viewer's affective response aligns with the narrator's framing rather than emerging from personal resonance with the image.

This process does not eliminate interpretation entirely. Viewers still respond to images, sounds, and narratives in individual ways. The range of acceptable interpretations, however, becomes narrower. Deviations from the intended meaning may feel less legitimate. The space for uncertainty diminishes, replaced by clarity and instruction.

The reduction of interpretive openness aligns

with the temporal constraints of short videos. Ambiguity requires time to resolve or reflect upon. In accelerated media environments, unresolved meaning risks disengagement. Voice-over resolves ambiguity quickly, ensuring that meaning is accessible within limited attention spans. Interpretive openness becomes a liability rather than a virtue.

Over time, this shift may influence broader habits of interpretation. Audiences repeatedly exposed to guided meaning may become less accustomed to extracting meaning independently from images. The skill of visual inference may weaken as narration assumes responsibility for explanation. Visual literacy practices adapt to a media environment where meaning is spoken rather than inferred.

### 5.3 Normalization of Guided Meaning

The widespread use of voice-over contributes to the normalization of guided meaning in short video culture. Guided meaning refers to a mode of communication in which interpretation is structured, directed, and stabilized through explicit cues. Voice-over is a central instrument of this guidance. It tells viewers how to understand what they see and why it matters.

This normalization is closely tied to platform dynamics. Short video platforms reward content that is easily understood and quickly consumed. Guided meaning reduces the risk of misunderstanding and increases the likelihood of engagement. Voice-over ensures that viewers grasp the intended message even during brief or distracted viewing. As a result, guided meaning becomes a dominant communicative strategy.

Standardization follows from this dominance. When many creators rely on similar narration styles, tones, and structures, meaning presentation becomes uniform. Videos across different topics and genres may share a similar explanatory rhythm. Introduction, explanation, conclusion are delivered through voice-over in predictable patterns. Visual diversity persists, but interpretive structure becomes standardized.

This standardization influences cultural expectations. Viewers learn to expect clear explanations and explicit framing. Content that requires interpretive effort may feel demanding or inefficient. Guided meaning becomes the norm against which other forms of communication are judged. Visual ambiguity or silence may be perceived as absence rather than invitation.

The normalization of guided meaning also intersects with authority. Voice-over often adopts an explanatory or instructive tone. The narrator positions themselves as knowledgeable, credible, or experienced. This establishes a hierarchy between speaker and viewer. Meaning flows from narrator to audience rather than emerging through shared exploration. Cultural authority is vested in the speaking voice.

This has implications for how knowledge and experience are mediated. Short videos increasingly resemble micro-lectures, commentaries, or explanations rather than visual narratives. Information is delivered rather than discovered. The culture of short video becomes one of instruction and guidance, even in entertainment-oriented content.

Guided meaning also affects the pace of cultural consumption. When interpretation is streamlined, content can be consumed more rapidly. Viewers move quickly from one video to the next without lingering uncertainty. This supports the endless scrolling logic of platforms. Voice-over plays a key role in sustaining this flow by minimizing friction in comprehension.

At the same time, guided meaning may limit the depth of engagement. When interpretation is predetermined, there is less incentive to reflect, revisit, or reinterpret. Content is consumed and moved past rather than contemplated. The cultural value of images as sites of exploration diminishes.

Taken together, the narrative and cultural implications of voice-over as an explanatory mechanism are substantial. The shift from showing to telling redefines storytelling norms. Reduced interpretive openness reshapes audience participation. The normalization of guided meaning aligns audiovisual communication with platform demands for speed and clarity. These changes do not signal the disappearance of visual culture but mark a transformation in how images function within contemporary media environments. Voice-over becomes not only a technical feature but a cultural force that reshapes narrative authority, interpretive practices, and the balance between sound and image.

## 6. Discussion

### 6.1 Voice-Over as an Adaptation to Visual Overload

The increasing centrality of voice-over in short videos should be understood as an adaptive

response to visual overload rather than a decline in creative standards or visual ambition. Contemporary media environments are saturated with images. Viewers are exposed to an uninterrupted stream of visual stimuli across platforms, applications, and interfaces. This saturation alters how images are perceived and processed. Individual images struggle to retain explanatory force when they appear within an endless sequence of competing visuals.

In such conditions, the capacity of images to sustain attention and generate meaning through observation alone is diminished. Visual explanation depends on the viewer's ability to slow down, compare, and integrate what is seen. Visual overload undermines this process by accelerating consumption and fragmenting attention. Voice-over emerges as a compensatory mechanism that restores coherence under these pressures. By providing a stable explanatory thread, narration allows meaning to persist even as images compete for attention.

This adaptation does not imply that visuals are unimportant. On the contrary, images remain essential for attracting attention and generating affective engagement. What changes is the division of labor between sensory channels. Visuals handle immediacy and emotional impact, while voice-over handles explanation and coherence. This division reflects a pragmatic response to conditions where images alone can no longer reliably perform both functions.

Voice-over also adapts to the temporal compression characteristic of short videos. Visual explanation often requires time to unfold, while narration can deliver meaning instantaneously. Under time pressure, voice-over becomes a rational solution. It allows creators to bypass the limitations of visual duration by stating what cannot be shown fully. The prevalence of voice-over should therefore be interpreted as a structural adjustment to accelerated media rhythms rather than a stylistic preference detached from context.

This perspective challenges critiques that frame narrated short videos as inherently simplistic or lazy. Such critiques often assume that visual explanation remains fully viable and that narration merely replaces effort. In reality, creators operate within constraints that make traditional visual explanation difficult to achieve. Voice-over represents an efficient and adaptive strategy within these constraints, not a

rejection of visual storytelling but a reconfiguration of its possibilities.

## 6.2 Rethinking Visual Dominance in Media Theory

The rise of voice-over as an explanatory substitute invites a reconsideration of visual dominance in media theory. Audiovisual studies have historically emphasized the primacy of images, often treating sound as secondary or supportive. This hierarchy reflects the influence of cinema studies, where visual composition, montage, and framing have been central analytical categories. Short video practices complicate this framework by foregrounding the explanatory power of voice.

In short videos, sound is not merely atmospheric or emotive. It is structural. Voice-over organizes narrative flow, establishes meaning, and guides interpretation. Without narration, many short videos would be difficult to understand despite their visual richness. This suggests that sound has assumed functions traditionally attributed to images. The explanatory center of gravity shifts from the visual track to the audio track.

This shift does not negate the importance of visual analysis but calls for a more balanced approach. Media theory must account for how sound and image interact under platform conditions that privilege speed, clarity, and accessibility. Voice-over demonstrates that meaning is not inherently visual in audiovisual media. Meaning emerges from the coordination of sensory channels, and the relative importance of each channel varies with context.

Rethinking visual dominance also requires reconsidering assumptions about spectatorship. Visual theories often presume attentive viewing and interpretive engagement. Short video consumption disrupts these assumptions. When attention is partial and fleeting, sound becomes a more reliable carrier of meaning. Media theory must therefore address not only textual properties but also viewing practices and technological environments.

The prominence of voice-over also highlights the role of language in shaping perception. Spoken narration does not simply describe visuals but frames how they are perceived. It influences what is noticed, how events are categorized, and which interpretations are privileged. This linguistic mediation of visual experience challenges the notion of images as transparent or self-evident carriers of meaning.



A revised theoretical framework would treat voice not as an accessory but as a central organizing principle in contemporary audiovisual media. Such a framework would examine how narration structures time, authority, and interpretation in platform-based environments. It would also explore how the balance between sound and image reflects broader cultural shifts toward guidance, efficiency, and explicitness.

### 6.3 *Limits of Voice-Over Substitution*

While voice-over effectively compensates for the limits of visual explanation, its dominance introduces new constraints and risks. One significant limitation concerns the expressive potential of visuals. When explanation is delegated primarily to speech, images may be relieved of explanatory responsibility but also deprived of narrative agency. Visuals become illustrative rather than exploratory. Their capacity to suggest, imply, or provoke reflection may be underutilized.

Excessive reliance on voice-over can lead to visual redundancy. Images may simply mirror what is being said rather than contribute independent meaning. This redundancy reduces the richness of audiovisual interaction. Instead of multiple channels offering complementary perspectives, meaning becomes centralized in speech. The audiovisual form risks collapsing into a spoken text accompanied by moving images.

Another limitation involves audience passivity. When narration delivers meaning directly, viewers are less required to interpret or infer. The cognitive work of meaning-making is shifted from the audience to the narrator. Over time, this may encourage habits of passive consumption. Viewers listen rather than observe, accept rather than explore. The interpretive skills associated with visual literacy may weaken as narration assumes explanatory authority.

Voice-over substitution also shapes power relations within media texts. The speaking voice often carries authority. It positions itself as knowledgeable and instructive. This can marginalize alternative interpretations and silence visual ambiguity. In contexts where narration frames images in a particular ideological or evaluative manner, viewers may have limited opportunity to question or reinterpret what they see. The voice becomes a

gatekeeper of meaning.

There are also aesthetic consequences. Silence, ambiguity, and visual rhythm have long been important expressive tools in audiovisual media. When voice-over becomes mandatory, these tools are constrained. Moments without narration may feel uncomfortable or incomplete within platform norms. The space for visual contemplation shrinks as continuous explanation fills every temporal gap.

These limitations do not imply that voice-over should be avoided. They suggest that substitution has costs as well as benefits. Voice-over resolves problems created by platform conditions but introduces new forms of constraint. Recognizing these trade-offs is essential for a nuanced understanding of contemporary media practices.

### 6.4 *Tensions Between Efficiency and Experience*

The discussion of voice-over as an explanatory mechanism reveals a broader tension between efficiency and experience in short video culture. Voice-over enhances efficiency by accelerating comprehension and reducing interpretive effort. It aligns with platform incentives that prioritize retention and engagement. At the same time, this efficiency may come at the expense of experiential depth.

Visual explanation often requires time and openness. It allows viewers to dwell on images, notice details, and construct meaning gradually. Voice-over compresses this process. Meaning is delivered rather than discovered. The experience of watching becomes more informational and less exploratory. This transformation reflects a cultural preference for speed and clarity but raises questions about what is lost in the process.

The tension between efficiency and experience is not easily resolved. Short video platforms are designed for rapid consumption. Expecting them to support slow visual explanation may be unrealistic. Voice-over responds effectively to these conditions. Yet acknowledging its limits opens space for critical reflection on platform design and cultural values.

Some creators experiment with hybrid approaches that balance narration and visual autonomy. They use voice-over sparingly, allowing images moments of silence or ambiguity. Such practices suggest that substitution is not inevitable but contingent. The

dominance of voice-over reflects prevailing conditions rather than fixed necessities.

Understanding this tension encourages a more reflective engagement with short video culture. Instead of celebrating or condemning voice-over outright, it becomes possible to analyze how it shapes experience, authority, and interpretation. Voice-over is neither inherently emancipatory nor inherently restrictive. Its effects depend on how it is used and the conditions under which it operates.

#### *6.5 Implications for Future Media Practices*

The discussion of voice-over substitution has implications for future media practices and research. As platforms continue to evolve, the balance between sound and image may shift again. New interfaces, attention patterns, or cultural preferences could alter the role of narration. Understanding current practices provides a foundation for anticipating such changes.

For creators, awareness of the limits of voice-over opens possibilities for experimentation. Narration can be used strategically rather than reflexively. Visuals can be given space to regain explanatory or expressive functions. Such choices may challenge platform norms but also expand the expressive range of short video.

For scholars, the prominence of voice-over invites renewed attention to sound, language, and voice in media analysis. Research can explore how narration shapes perception across different genres, cultures, and audiences. It can examine how voice-over interacts with issues of authority, identity, and power. The study of short videos thus becomes a site for rethinking fundamental assumptions about audiovisual communication.

The discussion underscores that voice-over as a substitute for visual explanation is both a solution and a transformation. It adapts audiovisual meaning-making to conditions of overload, acceleration, and partial attention. At the same time, it reshapes narrative authority, interpretive practices, and aesthetic possibilities. Understanding these dynamics requires moving beyond normative judgments and toward a contextual analysis of how media conditions shape the forms through which meaning is produced and received.

### **7. Conclusion**

This paper has examined the growing prominence of voice-over as an explanatory structure in short-form video and has argued that this development is inseparable from the conditions under which short videos are produced, distributed, and consumed. The central claim is that voice-over has not simply been added to short videos as a stylistic enhancement but has increasingly taken over the explanatory role once associated with visual continuity and visual demonstration. Fragmented images, accelerated editing, limited duration, and unstable attention environments have reduced the capacity of visuals to sustain meaning on their own. Spoken narration compensates for these limitations by organizing interpretation, summarizing complex relations, and maintaining coherence across discontinuous visual sequences.

The analysis has shown that visual explanation relies on temporal development, spatial coherence, and sustained attention, all of which are undermined by the structural logic of short video platforms. Under these conditions, images remain powerful as attention-grabbing and affective elements but lose reliability as explanatory devices. Voice-over emerges as a solution to this problem because language can condense information, clarify causality, and guide interpretation within severe time constraints. Meaning that once unfolded through visual observation is now delivered through speech.

The paper has also demonstrated that this shift is reinforced by broader media conditions. Everyday viewing practices are characterized by partial attention, multitasking, and mobile consumption. Sound remains perceptible in such contexts, while visuals are easily missed. Platform norms and algorithmic incentives reward content that communicates clearly and immediately, favoring narration as a stable and efficient channel. Over time, audiences have come to expect guided explanation, and creators have adapted their practices accordingly. Voice-over thus becomes normalized as an integral component of short video communication.

Taken together, these dynamics explain why voice-over functions not merely as support for images but as a substitute for visual explanation. Short videos increasingly rely on telling rather than showing, on verbal articulation rather than visual inference. This transformation reshapes

narrative structure, audience engagement, and the cultural role of images within platform-based media.

The primary theoretical contribution of this study lies in reframing voice-over as an explanatory mechanism rather than a secondary or decorative feature. Audiovisual media theory has often assumed the dominance of the image, treating sound and voice as supplementary layers that enhance visual meaning. The analysis presented here challenges this assumption by demonstrating that, in short video environments, explanatory authority frequently shifts from image to voice.

By conceptualizing voice-over as a substitute for visual explanation, the paper highlights a redistribution of explanatory labor across sensory channels. Meaning is no longer primarily produced through visual continuity or montage but through spoken narration that organizes perception and interpretation. This perspective invites a reconsideration of how audiovisual meaning is structured under platform conditions marked by acceleration, saturation, and attention scarcity.

The study also contributes to discussions of narration and perception by emphasizing the role of voice in guiding meaning. Voice-over does not simply convey information but shapes how images are seen and understood. It directs attention, frames emotional responses, and narrows interpretive possibilities. In doing so, it alters the relationship between text and audience, positioning viewers as recipients of guided meaning rather than co-constructors of interpretation.

This theoretical framing helps bridge media theory and platform studies. It situates aesthetic and narrative changes within the material and cultural conditions of digital platforms. Voice-over emerges not as an isolated stylistic trend but as an adaptive response to structural constraints. This approach avoids normative judgments that either celebrate or condemn narration and instead emphasizes contextual analysis.

More broadly, the paper contributes to a growing body of work that questions visual dominance in contemporary media. It suggests that sound, voice, and language play increasingly central roles in shaping audiovisual experience. Recognizing this shift opens new avenues for analyzing power, authority, and

meaning in platform-based communication.

The arguments presented in this paper point toward several directions for future research. One important area concerns genre differences. Short videos encompass a wide range of content types, including education, entertainment, advertising, personal storytelling, and political commentary. The function and prominence of voice-over may vary across these genres. Comparative studies could examine how explanatory strategies differ between instructional videos and narrative content or between commercial and user-generated media.

Cultural context represents another significant dimension. Platform practices are globally distributed but locally adapted. Expectations surrounding narration, visual clarity, and interpretive guidance may differ across linguistic and cultural settings. Cross-cultural research could explore whether the substitution of voice-over for visual explanation manifests similarly in different media cultures or whether alternative strategies emerge under different norms of communication.

Audience reception also warrants closer examination. While this paper has focused on structural and theoretical analysis, empirical research could investigate how viewers perceive and respond to narrated versus visually driven short videos. Studies could examine how voice-over affects comprehension, memory, emotional engagement, and trust. Such research would deepen understanding of how guided meaning shapes audience experience over time.

Another important direction concerns visual literacy. As narration increasingly carries explanatory responsibility, questions arise about the long-term effects on viewers' ability to interpret images independently. Research could explore whether habitual exposure to narrated media influences interpretive skills, tolerance for ambiguity, or engagement with non-narrated visual forms. This line of inquiry would connect short video studies to broader debates about media education and literacy. Future work could also address creative resistance and experimentation. While voice-over is dominant, some creators deliberately minimize or avoid narration, relying on visual storytelling despite platform constraints. Examining these practices could illuminate alternative possibilities within short video culture and reveal the limits of narration-centered explanation.

Voice-over as a substitute for visual explanation is not a marginal phenomenon but a defining feature of short video media. It reflects deep transformations in how meaning is produced, distributed, and consumed under platform conditions. Understanding this shift requires attention to narrative structure, perceptual habits, and cultural norms. By foregrounding voice-over as an explanatory mechanism, this study offers a framework for analyzing contemporary audiovisual communication and sets the stage for further inquiry into the evolving relationship between sound, image, and meaning.

## References

- Bordwell, D., & Thompson, K. (2019). *Film art: An introduction* (12th ed.). McGraw-Hill Education.
- Chion, M. (1994). *Audio-vision: Sound on screen* (C. Gorbman, Trans.). Columbia University Press.
- Chion, M. (1999). *The voice in cinema* (C. Gorbman, Trans.). Columbia University Press.
- Chovanec, J. (2020). Voice-over and presenter narration in TV documentaries. In J. Wildfeuer & J. Pflaeging (Eds.), *Media narrative theory* (pp. 155–170). Routledge.
- Kozloff, S. (1989). *Invisible storytellers: Voice-over narration in American fiction film*. University of California Press.
- Kruger, J.-L. (2010). Audio narration: Re-narrativising film. *Perspectives: Studies in Translatology*, 18(3), 231–249.
- Le Fèvre-Berthelot, A. (2013). Audio-visual: Disembodied voices in theory. *InMedia: The French Journal of Media Studies*, (3).
- Matamala, A. (2018). Voice-over: Practice, research and future prospects. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation* (pp. 465–480). Routledge.
- Pillai, S. E. (2015). The texture of interiority: Voiceover and visuals. *SVAC International Journal of Arts*, 2(1), 1–12.
- Smoodin, E. (1983). The image and the voice in the film with spoken narration. *Quarterly Review of Film Studies*, 8(1), 1–14.
- Vosmeer, M., Roth, C., & Koenitz, H. (2017). Who are you? Voice-over perspective in surround video. In G. Mark et al. (Eds.), *Interactive storytelling* (pp. 235–247). Springer.