ISSN: 3065-0003 | Volume 2, Issue 3

Open Access | PP: 43-48

DOI: https://doi.org/10.70315/uloap.ulirs.2025.0203007

Visual Storytelling Without Dialogue: Techniques of Emotional Framing in Short Video

Herasymiuk Heorhii

Videographer, Miami, USA.

Abstract

This article examines approaches to creating compelling visual storytelling without words in the format of short vertical videos on the TikTok and YouTube Shorts platforms. The study sought to identify key micro-tools and to correlate their effectiveness with the algorithmic requirements of modern platforms. The relevance of this work is justified by the rapid growth of TikTok and YouTube Shorts against the backdrop of the dominance of smartphone portrait orientation and the expanding practice of viewing video without sound. The high-speed environment of algorithmically personalized feeds requires authors to condense their narrative into a series of visually and emotionally engaging signals that are recognized by the viewer within the first few seconds and remain comprehensible regardless of linguistic or audio context. The novelty of the research lies in its interdisciplinary approach, which combines neurocognitive data on the functioning of mirror neurons and embodied simulation, a meta-analysis of documentation from TikTok and YouTube platforms, quantitative engagement metrics, and eye-tracking studies of editing patterns. A unified comparative analysis framework is proposed, linking neurophysiological mechanisms, the algorithmic constraints of recommendation systems, and the practices of short-video creators. The main findings demonstrate that an instantaneous compositional hook activates motor-mirror resonance, which is then imbued with affective valence by color and light contrast. That rhythmic editing, combined with geometric camera movement, produces a singular impulse that holds the viewer's attention within the 60-second constraint. The synergy of these elements ensures the universality of wordless visual storytelling and promotes maximum organic distribution of content in recommendation feeds. This article will be beneficial to media and marketing researchers, digital content specialists, and creators of short videos.

Keywords: *Emotional Framing; Visual Storytelling; Short Vertical Video; TikTok; YouTube Shorts; Neurocognitive Mechanisms; Algorithmic Distribution; Rhythmic Editing.*

INTRODUCTION

During the 2020s, short vertical videos have evolved from an experimental format into the central domain of mobile viewing. TikTok reached 1.59 billion monthly users [1], and YouTube Shorts—launched only in 2021—registered approximately 200 billion views per day by summer 2025 [2]. Simultaneously, smartphones are held in portrait orientation 91% of the time, and vertical clips exhibit up to a ninefold increase in completed views compared to their horizontal counterparts, making them the most natural medium of the endless-scroll era [3].

This expansion is driven not only by demographic reach but also by the specifics of algorithmic distribution: platforms present short clips in a personalized, recommendation-based skimming feed, where the interval between swipes is measured in mere hundreds of milliseconds. Under these conditions, quantitative metrics (watch-through rate, rewatch rate, CTR) depend directly on the moment a clip succeeds in evoking an emotional response. The vertical frame minimizes viewers' cognitive load by instantly filling the screen and eliminating the noise margins characteristic of horizontal video on mobile devices.

However, the accelerated visual environment has given rise to a paradox: a significant portion of the audience watches content with the sound muted. A Sprout Social study shows that 74% of social media video views occur in mute mode [4]. The absence of a verbal track thus becomes not a limitation but a structural challenge: creators are compelled to condense narrative into a sequence of visually emotional signals that must be decoded within the first seconds and remain intelligible regardless of the language environment or device volume level.

Citation: Herasymiuk Heorhii, "Visual Storytelling Without Dialogue: Techniques of Emotional Framing in Short Video", Universal Library of Innovative Research and Studies, 2025; 2(3): 43-48. DOI: https://doi.org/10.70315/uloap.ulirs.2025.0203007.

MATERIALS AND METHODOLOGY

The study draws on 17 sources encompassing platform analytics, neurocognitive research, editing practices, and the platforms' guidelines. Its empirical foundation comprises TikTok and YouTube Shorts usage statistics [1, 2], data on smartphone portrait orientation usage [3], and social media silent-viewing figures [4]. The theoretical framework is established by works on mirror-neuron mechanisms in action perception [5, 6] and analyses of user engagement in Shorts versus regular videos [7]. The algorithmic context is reconstructed through studies of TikTok's and YouTube's recommendation systems, highlighting the weight of completion and emotional valence in ranking [8, 9], as well as through the platforms' official creative guidelines [10, 11]. Additionally, quantitative assessments of the impact of audiovisual features on engagement in Douyin and Shorts are incorporated [12, 15].

Methodologically, the work combines several approaches. First, a systematic review and comparison of platform documentation: TikTok's requirements for a compositional hook and its critical retention windows (3–6 s) [10, 11] are set against the gradient-boosting structures of YouTube Shorts [9] and the weightings of the multivalent For You Feed model [8]. Second, a meta-analysis of neuroimaging and cine-psychological experiments—evaluating the onset time of embodied simulation (within the first second) and its relation to premotor cortex activation [5, 6].

For quantitative analysis, statistics on video duration and engagement were utilized, including the growth in the share of videos up to 60 seconds in Shorts during 2021–2023 [7], optimal average shot durations and full completion rates [15], as well as eye-tracking data on editing patterns [16]. To verify color and lighting techniques, an experiment with 117 respondents was considered, measuring valence changes under monochrome versus color conditions [12]. Finally, a content analysis of user scenarios drew on scrollytelling research [17], examining engagement mechanics via dynamic layer shifts and text anchors. All data were processed within a unified comparative analysis framework, enabling the integration of neurocognitive mechanisms, algorithmic constraints, and the practices of short vertical video creators.

RESULTS AND DISCUSSION

Cognitive film studies over the past two decades have demonstrated that the viewer first decodes the emotional meaning of a shot. When observing an action, the brain activates mirror neurons, thereby trying on another's experience within one's bodily encoding [5]. The result of the study is shown in Figure 1.

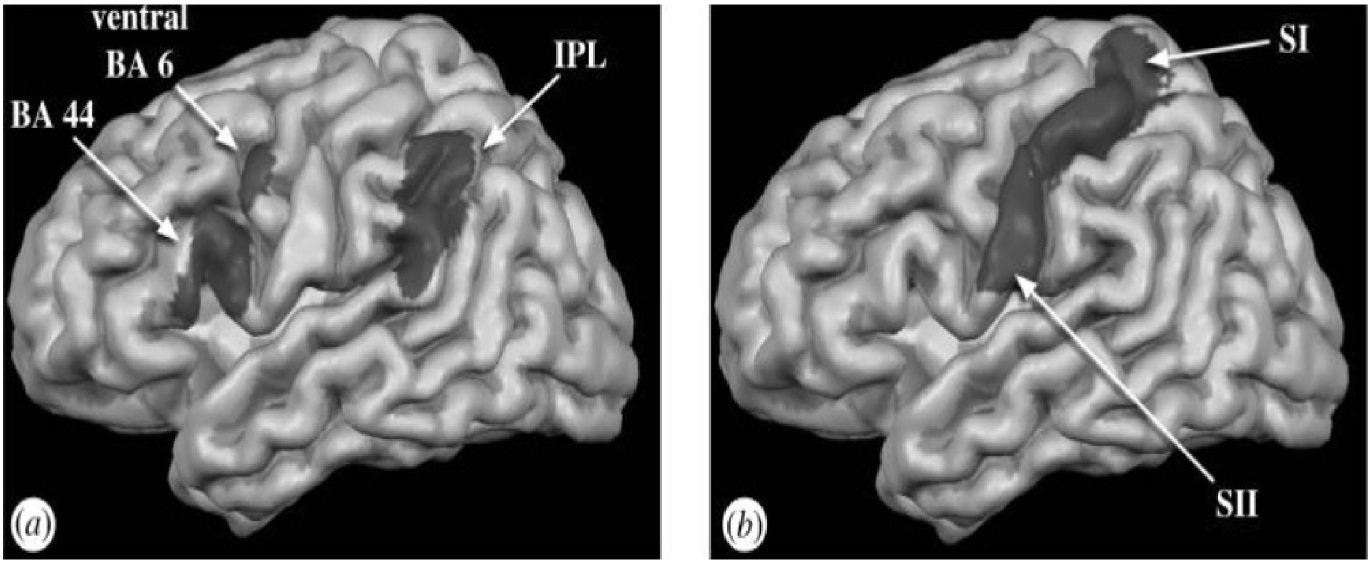

Fig. 1. Anatomical locations of the motor and somatosensory components of simulation. (a) Lateral view of the human brain with the location of the ventral premotor cortex (BA6/BA44) and the inferior parietal lobule (IPL). (b) Lateral view showing the location of the primary and secondary somatosensory cortex (SI/SII) [6]

Meta-analyses of neuroimaging studies report a correlation between the intensity of premotor cortex activity and subjective empathy ratings when viewing dynamic scenes, a finding that holds for both artistic representations and real actions. For short vertical videos, this implies that the maximum effect must be delivered instantaneously: the faster the viewer recognizes a motor program familiar to them in a character's or object's movement, the higher the likelihood of watch-through and replay [6].

The compression of the temporal range to 60 seconds compels creators to adopt micro-dramaturgy — a narrative form in which the classical acts of exposition, conflict, and resolution collapse into a single coherent gesture [7]. Quantitative YouTube analyses indicate that between 2021 and 2023, the proportion of videos up to one minute long published as Shorts more than doubled, as shown in Figure 2.

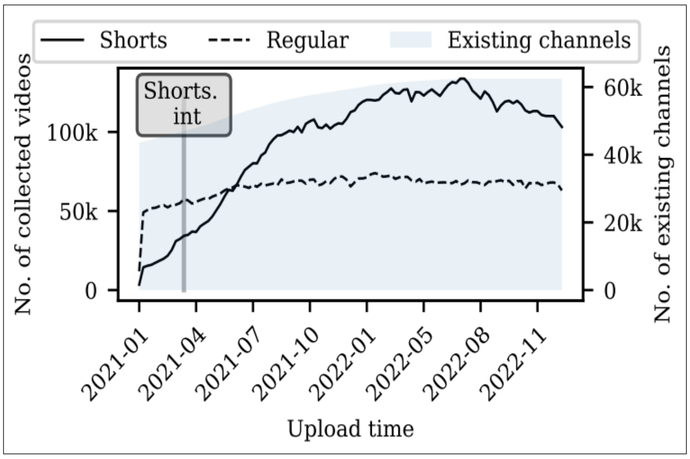

Fig. 2. Weekly video uploads, categorized into Shorts and RVs, with the number of channels older than the respective week [7]

Thus, limited time does not dismantle narrative but translates it into a mode of singular impulse: a single visual hook serves both as setup and conflict, while the climactic edit serves as resolution.

This compressed form is reinforced by platform architecture. TikTok's For You Feed algorithm ranks clips using a multi-signal model in which the weight of a completed view exceeds that of a like or comment, and replay is marked as an indicator of high content satisfaction [8]. YouTube Shorts' algorithm, trained via watch-time-optimized gradient boosting, additionally factors in emotional valence, as demonstrated by experiments manipulating the tone of recommended content: recommendations amplify users' already demonstrated affective preferences [9]. In such an environment, every second of ambiguity translates into a swipe risk, and all intermediate frames function only to maintain the kinetic pulse that sustains mirror-simulation activity.

The combination of neurophysiological resonance, compressed narrative, and algorithmic selection forms a new grammar of short video, wherein an expressive gesture or rhythmic transition between shots performs the same role as a character's line in classical drama. The creator thus operates simultaneously on three axes: designing the shot to instantaneously trigger embodied simulation, structuring the plot as a unified syntactic unit, and synchronizing this solution with the metrics embedded in recommendation models. This triple constraint does not diminish expressiveness; instead, it fosters the invention of new forms of emotional framing that are universally intelligible to any viewer, regardless of interface language or whether sound is enabled.

Emotional framing in short vertical video begins with the so-called compositional hook — an optical event that the viewer decodes before consciously recognizing what they have seen. The platform itself enforces this requirement: TikTok's advertiser help centre underscores that the core message must be delivered within the first three seconds; otherwise, the watch-through probability falls exponentially, and brand recall is scarcely formed [10]. TikTok's internal marketing playbook further refines the 3–6 s range as the critical window for attention retention and cites a survey in which 79% of users agreed that they expect the creator's or brand's personality to manifest at that moment [11]. From a neurocognitive perspective, this technique is effective because the sudden appearance of a hand, face, or dynamic object instantly activates mirror neurons in the premotor cortex, creating an internal sense of participation in the action rather than merely observing it externally. Practically, this translates into a rule: the face or symbolic object should be placed in the upper third of the 9:16 frame, where it lies closest to the swipe-finger position and intercepts the motor intention to scroll the feed.

The second axis of emotional framing is the contrast of color and light. Neuro- and psychophysiological studies confirm that a change in color palette alone can shift perceptual valence: in an experiment with 117 respondents, the transition from monochrome to color reliably increased their ratings of a neutral face's joyfulness, whereas the combination of cool tones with frightening editing intensified the feeling of anxiety, as recorded both on behavioral scales and in ACC and insular activity [12], as shown in Figure 3.

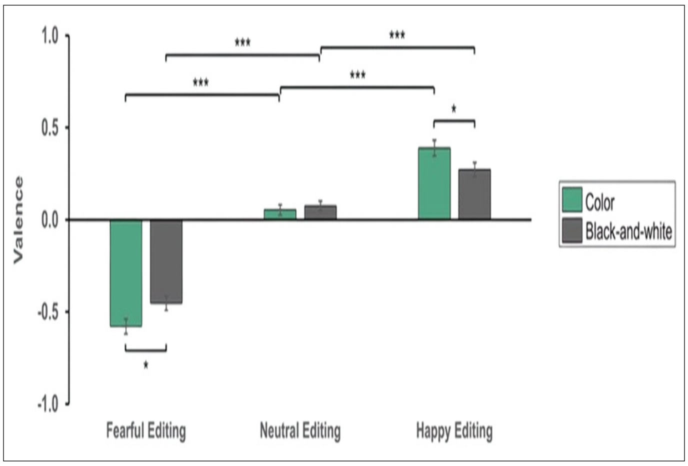

Fig. 3. Neutral Face Valence in Color and Grayscale across Laptop and MRI Contexts [12]

For short-form video, this implies that color must be embedded not as a decorative layer but as an autonomous dramaturgical stimulus: a flash of warm orange against a cool blue background instantly signals an emotional shift. It prepares the viewer for a narrative turn.

More applied insights emerge from the analysis of 2,553 Douyin advertising clips: brightness and color saturation proved significant predictors of the number of likes and comments, whereas dark, low-contrast clips—despite identical duration—elicited fewer interactions; the authors attribute this to the fact that increased brightness facilitates visual feature extraction on the small screen [12]. Consequently, it is recommended to introduce a striking color accent within the first third of a clip's duration and to add further light contrast toward the climax, thereby marking the emotional peak without relying on verbal explanation.

Thus, the compositional hook and the contrast of color/light operate synergistically: the former immediately triggers empathic resonance via the motor-mirror system, while the latter converts that resonance into a specific affective valence. Their joint application enables a full-fledged wordless narrative within the 60-second constraint: the viewer not only comprehends what is occurring but also feels how they should respond affectively, and the platform registers this in increased full-view rates.

In short vertical videos, the editing rhythm performs the same function as the compositional hook and the color accent—capturing the viewer before the cognitive filter can initiate a swipe. In classical editing terminology, one distinguishes a metric cut, in which shots change at fixed frame-count or beat intervals, from a rhythmic cut, in which shot durations adapt to the internal movement of the subject or to music. Early editing manuals emphasized that the metric technique creates a sense of inexorable tempo. In contrast, the rhythmic approach allows the scene's emotion to breathe, combining frame speed with the fluidity of action [13]. In the context of TikTok and YouTube Shorts, this dichotomy is compressed: the platform rewards a narrow metric range between cuts, yet within each impulse, creators embed a micro-rhythm of movement, allowing the viewer to experience natural dynamism rather than mechanical fragmentation.

Neuropsychological experiments demonstrate that rapid shot changes exaggerate the subjective duration of an episode; scenes edited with continuous cuts are perceived as longer than their actual runtime, thus holding attention longer than static scenes of equal duration [14]. An additional analysis of 100 popular Shorts from 2024 records an average shot length of approximately 2.5 seconds; clips with this tempo achieve 35% more full views than those with shots longer than four seconds [15]. Contemporary short-form video effectively combines a metric framework with rhythmic modulation: the overall structure is defined by uniform impulses, yet within them, there is variable motion synchronized to the beat or the protagonist's gesture, reducing cognitive fatigue and stimulating a dopamine response to novelty.

Editing tempo is directly tied to gaze trajectory: an eye-tracking study of film sequences showed that fixation points after each cut predictably migrate to the center and then follow movement lines within the frame; this attention pull is stronger when editing is supported by an internal motion motive rather than by mere uniform alternation of shots [16]. The camera, employing the same logic, expands emotional range through geometric and parallax shifts. A diagonal tracking shot, circular tracking, or frontal zoom-in not only imbues the scene with kinetic energy but also manipulates depth of field, creating in the viewer a sense of immersion in the spatial environment.

In practice, this means that a parallax move should be deployed where one needs to amplify emotional amplitude without words. For example, a rapid approach to a symbolic object at the climactic moment allows one to visually emphasize the resolution without resorting to text or speech. User studies in the scrollytelling genre further confirm that dynamic layer shifts enhance subjective engagement compared to static long-form layouts, and this principle transfers readily to the mobile vertical screen [17].

Thus, rhythmic editing and geometric camera movement form an interconnected mechanism: the former establishes a temporal pulse, the latter transforms it into a spatial attraction. Together with the compositional hook and the color contrast discussed above, they constitute a quartet of micro-tools capable of triggering the mirror-neuron response and retaining the viewer within the sixty-second limit, without a single word being uttered.

The practical logic of short vertical video begins at the moment the viewer's finger hovers over the screen: the first seconds determine whether they remain engaged with the clip or continue swiping. Therefore, the compositional hook described above must not merely appear quickly but must immediately reveal a micro-conflict: a gesture, a rapid zoom, or a sharp color accent forms an impulse that instantaneously activates motor and affective mirroring. Should the creator permit any preparatory build-up, the algorithm interprets the ensuing drop in engagement as a signal of low quality and truncates the clip's organic reach.

The second layer of strategy involves vertical framing, which accounts for the unchanging interface mask. The avatar and description overlay the lower portion of the screen, the like counter occupies the right edge, and the audio waveform bar appears in the upper-left corner. Key elements of the action should be positioned along a diagonal corridor from the upper third to the center of the frame; here, they remain visible regardless of the phone model, and the viewer can interact with the interface without obscuring the narrative. Simultaneously, it is crucial to distribute dynamics by depth: a symbolic object may be brought to the foreground while the background remains relatively static—thus the viewer's gaze more readily fixes on the semantic core.

Since a significant portion of views occur without sound, textual captions become an anchor that retains attention and establishes context. Mini-subtitles of up to one line serve two functions: they convey meaning if speech is present, and they set an anticipatory vector if the clip is entirely silent. These are placed in an area free of interface icons, typically near the upper or middle third of the screen. The typographic solution must be high-contrast yet unobtrusive, so as not to compete with the visual impulse: a brief trigger word or question encourages the viewer to watch through the video to confirm or refute the assumption raised by the text.

Wordless narratives benefit from a dichotomous before/after plot, especially in transformation, repair, or personal makeover formats. A single visual snap, such as an abrupt cut on a clap or a shift in angle, transitions the scene from its initial state to its final one, simultaneously serving as both the climax and the resolution. This device is intuitively comprehensible to any audience: the cognitive system recognizes the contrast and reconstructs the causal link without requiring verbal explanation. The more pronounced the contrast between the two states: in form, color, scale, or emotional expressiveness, the more vividly the effect of change is experienced.

Finally, optimal publication timing takes into account algorithmic distribution triggers. Recommendation technologies track full views, replays, and interactions in the clip's closing phase. To stimulate cyclical replay, the final frame is edited so that visual and temporal seams coincide: the action concludes at a point that logically allows a return to the beginning, creating a seamless loop. Simultaneously, a call to action, such as commenting, liking, or transitioning to the next video in a series, is engaged. The CTA must be integrated into the frame itself, rather than presented as a separate title card. The protagonist's gesture toward the lens or continuation of physical movement beyond the frame cues the viewer that active participation is the natural extension of viewing.

Thus, when all the aforementioned elements—early hook, proficient vertical composition, concise text, contrasting dual-state narrative, and seamless loop with unobtrusive CTA—are assembled into a unified rhythmic-spatial structure, the clip meets both cognitive and algorithmic requirements simultaneously, achieving maximal organic distribution without reliance on dialogue.

CONCLUSION

The conducted research reveals that unconditional emotional engagement in short vertical video is formed at the intersection of neurocognitive mechanisms and platform algorithmic constraints. Activation of mirror neurons via an instantaneous compositional hook ensures an immediate response, reinforced by color and light contrast as autonomous dramaturgical stimuli. The 60-second temporal compression compels creators to employ micro-dramaturgy: all classical narrative elements—exposition, conflict, and resolution—are condensed into a single, visually emotional gesture that can be decoded regardless of interface language or sound availability.

Concurrently, the platform prioritizes its metrics, where completed views and replays serve as key quality indicators. TikTok's and YouTube Shorts' ranking algorithms incentivize the optimization of every shot and every transition according to engagement metrics. Early visual impulse, rhythmic editing, and geometric camera techniques operate in synchrony to minimize swipe risk and maximize attention retention time. Moreover, structuring the frame space around interface elements and employing concise textual anchors further enhances the visibility and clarity of the core action.

Accordingly, wordless, short vertical videos establish a new universal grammar of emotional storytelling. The quartet of micro-tools—compositional hook, color contrast, rhythmic editing, and geometric camera movement—merges with algorithmic triggers and interface constraints to form a single rhythmic-spatial framework. This synergy not only delivers a complete emotional and narrative experience within one minute but also adapts most effectively to the content-distribution principles of contemporary platforms.

REFERENCES

1. S. Singh, "How Many People Use TikTok," Demand Sage, Jan. 01, 2025. https://www.demandsage.com/tiktok-user-statistics/ (accessed May 25, 2025).
2. K. Cobb, "YouTube Shorts Now Averages 200 Billion Daily Views," TheWrap, 2025. https://www.thewrap.com/youtube-shorts-200-billion-daily-views/ (accessed May 26, 2025).
3. "Smartphone vs Tablet Orientation: Who's Using What?" Scientia Mobile, 2018. https://scientiamobile.com/smartphone-vs-tablet-orientation-whos-using-what/ (accessed May 27, 2025).
4. A. Ahmed, "Social Media Video Statistics Marketers Need to Know," Sprout Social, Mar. 26, 2024. https://sproutsocial.com/insights/social-media-video-statistics/ (accessed May 28, 2025).
5. Z. Clay and M. Iacoboni, "Mirroring Fictional Others," Emory, Dec. 2011. Accessed: May 29, 2025. [Online]. Available: https://www.emory.edu/LIVING_LINKS/publications/articles/Clay_Iacoboni_2011.pdf
6. J. A. C. J. Bastiaansen, M. Thioux, and C. Keysers, "Evidence for mirror systems in emotions," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 364, no. 1528, pp. 2391–2404, Aug. 2009, doi: https://doi.org/10.1098/rstb.2009.0058.
7. C. Violot, T. Elmas, I. Bilogrevic, and M. Humbert, "Shorts vs. Regular Videos on YouTube: A Comparative Analysis of User Engagement and Content Creation Trends," arXiv (Cornell University), Mar. 2024, doi: https://doi.org/10.1145/3614419.3644023.
8. T. Annabell, "TikTok Search Recommendations: Governance and Research Challenges," Arxiv, 2024. https://arxiv.org/html/2505.08385v1 (accessed May 30, 2025).
9. H. Habib, "YouTube Recommendations Reinforce Negative Emotions," Arxiv, 2019. https://arxiv.org/html/2501.15048v1 (accessed Jun. 01, 2025).
10. "Creative best practices for performance ads," TikTok, 2024. https://ads.tiktok.com/help/article/creative-best-practices?lang=en (accessed Jun. 02, 2025).
11. TikTok, "Best practices and tools to make your ads a smash hit," TikTok, 2023. Accessed: Jun. 04, 2025. [Online]. Available: https://ads.tiktok.com/business/library/SMB_Creative_Playbook_External.pdf

12. C. Wang and Z. Li, “Unraveling the relationship between audience engagement and audiovisual characteristics of automotive green advertising on Chinese TikTok (Douyin),” PloS one, vol. 19, no. 4, pp. e0299496–e0299496, Apr. 2024, doi: https://doi.org/10.1371/journal.pone.0299496.

13. E. Dmytryk, On Film Editing. Routledge, 2018. doi: https://doi.org/10.4324/9780429506086.

14. K. Kovarski, J. Dos Reis, C. Chevais, A. Hamel, D. Makowski, and M. Sperduti, “Movie editing influences spectators’ time perception,” Scientific Reports, vol. 12, no. 1, Nov. 2022, doi: https://doi.org/10.1038/s41598-022-23992-2.

15. M. D. Cena, “Video Clip Length: Ultimate Guide for Every Editing Style,” Vidpros, 2025. https://vidpros.com/video-clip-length/ (accessed Jun. 07, 2025).

16. A. Bruckert, M. Christie, and M. O. Le, “Where to look at the movies: Analyzing visual attention to understand movie editing,” arXiv, Jan. 2021, doi: https://doi.org/10.48550/arxiv.2102.13378.

17. A. Tjärnhage, U. Söderström, O. Norberg, M. Andersson, and T. Mejtoft, “The Impact of Scrollytelling on the Reading Experience of Long-Form Journalism,” Proceedings of European Conference in Cognitive Ergonomics 2023, Sep. 2023, doi: https://doi.org/10.1145/3605655.3605683.