

Reference Fidelity in AI-Generated Short-Form Video Across Four Prompt Packages

Chloe Zihan Jin

*Orange County School of the Arts, Santa Ana, USA
ocsa.administration@ocsarts.net*

Abstract. Generative video systems can produce short clips from textual and visual instructions, yet their ability to preserve the content of a human reference remains uncertain. This exploratory study examines 160 videos generated with Doubao-Seedance-2.0 from 40 human-made short-form videos drawn from Douyin and Xiaohongshu. The references were divided evenly among humor, informational, demonstration, and storytelling content. Qwen3.8-Max analyzed each reference and produced four prompt packages labeled vague, vivid, keyword, and image-based. The fourth package also contained timestamped reference frames, frame descriptions, identity constraints, and sequence instructions, so it is treated as a multimodal reference condition rather than a pure image condition. Each prompt was used once under the same default settings, and no output was edited after generation. Metadata inspection covered the complete corpus. All 40 groups received a standardized frame review, and eight groups were examined at three temporal positions. Outputs usually retained the source topic and central action. Fidelity was stronger for one-subject situations than for multi-step procedures, factual labels, comic timing, or longer causal sequences. The multimodal package appeared closer to source composition in several cases, yet the one-shot design and unequal prompt information prevent a causal ranking. The study therefore evaluates reference fidelity rather than virality or prompt efficacy.

Keywords: generative AI, short-form video, prompt packages, reference fidelity, human-AI co-creation

1. Introduction

Short-form video combines moving images, written language, speech, music, and event sequence within a compressed format. A recent multimodal annotation study argues that these layers may operate asynchronously and therefore require separate analytical units [1]. Generative systems add another layer of mediation because a creator specifies a desired output and a model chooses subjects, actions, camera positions, and temporal transitions. Lubart describes the computer-as-colleague model as a cycle in which a human proposes an idea, the computer modifies it, and the human selects or refines the result [2]. The COFI framework similarly treats human-AI co-creation as interaction around a shared creative product and identifies participation, contribution, and communication as central components [3].

Prompting is part of that interaction, although prompt wording alone does not provide complete control. A CHI study of non-expert prompt designers found repeated local testing, overgeneralization from limited examples, and difficulty diagnosing failures [4]. The current experiment avoids iterative prompting. Every prompt was used once, which creates a record of first-attempt output while leaving model randomness unmeasured. The four conditions also differ in both format and information quantity. The experiment therefore compares prompt packages rather than isolating the effect of a single variable such as descriptive length or image input.

This paper explores how faithfully AI-generated videos preserve the core visual, textual, and narrative features of human reference clips in terms of topics, subjects, actions, composition, visible text and event progression. It further investigates fidelity variations across video outputs induced by vague, vivid, keyword-based and multimodal prompt packages. The findings contribute a framework for assessing reference fidelity and inform the design of more effective prompting strategies for AI-generated video.

2. Literature review

2.1. Human-AI creative partnership

Creative partnership requires the human participant to communicate goals and judge the system's contribution. Lubart's colleague model assigns computers a role in generating, evaluating, or refining ideas while retaining human selection and development [2]. COFI separates interaction between collaborators from interaction with the shared product and includes turn-taking, contribution type, communication, and editing [3]. These concepts fit generated video because a prompt leaves many production decisions unspecified. The output records both the supplied constraints and the model's interpretation.

Prompt comparison also requires careful experimental language. Zamfirescu-Pereira and colleagues studied natural-language prompt design for an LLM-based chatbot and found that participants often relied on repeated local trials and generalized from limited observations [4]. Their result does not establish rules for video prompting. It does show why a single output should not be treated as a stable estimate of a prompt's performance. In the present study, prompt packages vary in wording, length, and modality. Observed differences describe these complete packages and cannot identify the independent contribution of any one component.

2.2. Features of viral and popular short-form video

Virality and popularity require separate evidence. Berger and Milkman found that high-arousal emotions, practical utility, interest, and surprise were associated with New York Times sharing, while their experiments tested the effect of specific emotions on transmission [5]. Their unit of analysis was mainly written news content, so the findings are contextual background rather than a checklist for short videos. Broxton and colleagues classified video referrals as social or non-social and found that popular videos were not always highly social [6]. France, Vaghefi, and Zhao modeled initial viral, delayed viral, and other viewing patterns from longitudinal view data [7]. These studies show why a static file cannot verify viral behavior.

Video studies provide a second caution. Jiang and colleagues built a viral-video benchmark and predicted peak timing from view-count sequences and metadata [8]. A TikTok study of 400 videos found creator follower count to be its strongest predictor, with shot scale, text presence, point of view, and video lifespan also contributing in that sample [9]. A study of more than 260,000 Douyin

videos found a strong correlation between views and likes, while other popularity metrics had weaker relationships [10]. Content features matter within these studies, yet distribution history, creator status, and the chosen outcome measure also matter.

The present review uses content dimensions without converting them into a virality score. The short-video annotation framework developed by Grzenkiewicz and Wildfeuer separates shots, visible objects, written language, sound, speech, and event structure because these elements may begin and end at different moments [1]. Following that logic, this study examines topic, focal subject, action, composition, visible text, and event progression as distinct aspects of reference correspondence. Audio is identified in metadata but is not transcribed or rated.

3. Methods

3.1. Corpus and prompting conditions

The corpus contained 40 human-made source clips collected from Douyin and Xiaohongshu and selected through content-type and keyword screening. Ten clips were assigned to each of four operational genres. Humor clips centered on comic behavior or incongruity. Informational clips explained factual material. Demonstration clips showed actions or procedures. Storytelling clips contained characters and an event sequence with an opening and an ending. The accompanying category list was a screening guide rather than a literal inventory of the selected clips. The source links, account identifiers, publication dates, and audience metrics were not supplied. The clips are therefore described as references selected for genre fit.

Qwen3.8-Max analyzed every reference video and produced four English prompt packages. M1 was a short summary labeled Vague Prompt. M2 was a long scene description labeled Vivid Prompt. M3 was a list of visual and semantic terms labeled Keyword Prompt. M4 was labeled Image-Based Prompt and contained several timestamped reference frames, prose descriptions of those frames, instructions to retain characters and settings, and an ordered account of the intended action. Because M4 combined visual evidence with detailed textual constraints, this paper calls it the multimodal reference condition. The complete prompt catalogue contains 40 instances of each package, for 160 prompts in total.

Each prompt was submitted once to Doubao-Seedance-2.0. All four conditions used the same default settings. The supplied record does not enumerate those defaults, the precise model build, the seed, or the generation date. No output was edited, cropped, retimed, subtitled, or given additional music after generation. The design produced one reference and four generated outputs per group, for 200 MP4 files overall. Since there was one output per prompt, no selection among repeated attempts occurred. The design also cannot separate prompt-package effects from stochastic generation.

A batch metadata inspection preceded visual comparison. The 40 references ranged from 8.367 to 70.767 seconds and had a mean duration of 22.400 seconds. Their resolutions and orientations varied. All 160 generated clips used a 560 by 752 vertical frame at 24 frames per second. A total of 148 outputs lasted approximately 10.1 seconds, while 12 lasted approximately 15.1 seconds. Every file contained an audio stream. The uniform generated format and shorter duration created a structural difference from many references.

3.2. Comparative review procedure

The review used two stages. First, file correspondence and technical metadata were checked for the complete corpus. A midpoint frame was extracted from every file and arranged by matched group,

allowing all 40 groups to be screened under the same visual procedure. Second, eight groups were selected for closer analysis, with two from each genre. Within every genre, one example represented a relatively simple action or stable situation and one represented a more complex procedure or multi-stage idea. The detailed sample therefore contained eight reference clips and 32 generated clips.

For each video in the detailed sample, frames at approximately 20 percent, 50 percent, and 80 percent of the duration were compared. The review recorded central topic, major subjects, actions, composition, setting, visible text, and apparent event progression. These categories reflect the separation of visual, textual, and event layers recommended in multimodal short-video analysis [1]. Observations were descriptive. No numeric scale, inferential statistics, or formal inter-rater reliability calculation was used. Audio streams were not transcribed or scored.

4. Results

4.1. Overall correspondence

Most generated clips in the midpoint screening retained the broad semantic themes of their references, such as keeping a dog central in dog interactions and educational examples remaining instructive. However, exact replication was less consistent. Generated clips often replaced people, settings, props, camera positions, and action order. Several clips looked coherent independently while departing from the source procedure or narrative.

The difference between simple and complex references was visible across genres. Stable situations with a single relationship were easier to preserve. In the storytelling example of a child reading beside a cat, all four conditions were retained: the girl, book, cat, and calm emotional situation. In a product demonstration involving a handheld fan, all conditions preserved the presenter and demonstration structure, although one output showed an unclear device at an early point. By comparison, the coded keypad and safe demonstration produced substantial changes in props and causality. Some variants showed keys or a safe, while another introduced destruction or an explosion. The clips retained the idea of a coded device but did not preserve a reliable procedure.

4.2. Genre-specific observations

Humor outputs generally reproduced recognizable setups while changing the details that could affect comic timing. In the card-guessing example, the generated versions retained a card near a participant's forehead and a group reaction, yet the cast, gestures, and implied sequence differed. In the dog-size contrast example, all conditions preserved an encounter between dogs, while breeds, location, and the number of animals varied. One condition introduced several small dogs, which changed the visual premise. The multimodal reference version was closer to a two-dog confrontation in this case.

Informational outputs exposed a separate problem involving factual text. In the artery and capillary example, several variants retained a staged educational presentation and anatomical labels. One vivid-prompt output displayed the misspelt label Artern, reducing factual reliability even though the overall visual concept remained clear. In the example concerning the historical development of swings, the outputs covered related ideas such as swinging, primates, tools, and historical settings, but their sequences diverged. The multimodal reference version showed a clearer multi-stage structure and included Chinese captions. The result was thematically relevant without being a close reproduction of the reference progression.

Storytelling clips frequently preserved emotional relationships more effectively than exact event order. In an animal rescue example, the generated versions retained dogs, cats, protection, or care in different combinations. The M4 sequence appeared comparatively coherent and showed a dog caring for cats, while other conditions altered the initial setup or the rescue mechanism. These observations suggest that the system can maintain a broad emotional direction while inventing the actions used to express it.

4.3. Prompting condition patterns

No prompt package was uniformly superior across the reviewed cases. M1, M2, and M3 often preserved the topic while varying the cast, setting, props, and event order. Greater verbal detail did not prevent text errors or procedural changes. M4 appeared closer to source composition or subject relationships in several reviewed examples, including the dog encounter, fan close-up, swing sequence, and animal rescue. M4 also carried more source information than the other packages. It included multiple frames, timestamps, frame descriptions, identity constraints, and sequence directions. The observed proximity cannot therefore be attributed to image input alone. Other M4 outputs still introduced new information. The result is a descriptive pattern across unequal prompt packages rather than evidence of a causal hierarchy.

5. Discussion

5.1. Prompt information and creative control

The findings fit a view of generative video as human-AI co-creation. The creator supplies a partial specification, and the system contributes visual and temporal decisions [2, 3]. In this experiment, a second model also mediated the instruction because Qwen3.8-Max converted each source clip into the four prompts. The evaluated pipeline therefore includes source selection, machine analysis, prompt construction, and video generation. Departures may enter at either model stage. The outputs usually retained the broad subject, while control weakened when execution depended on exact order, a specific prop, factual text, or a precise setup and payoff.

The full prompts change the interpretation of M4. Its frames supplied spatial evidence, while its prose supplied action order, identities, settings, caption style, and temporal progression. A closer M4 result could arise from any combination of those elements. M2 was also substantially longer than M1, and M3 changed sentence structure into a keyword list. A controlled follow-up would hold information content constant while varying one feature, repeat each prompt with recorded seeds, and score outputs with independent raters. The present corpus instead documents how four naturally constructed instruction packages behaved on their first generation.

5.2. Relevance to features associated with viral video

The generated clips used a consistent vertical frame and mostly lasted about ten seconds. Many established a subject quickly and retained a recognizable genre. These are observable formal properties, rather than evidence of success. Research on sharing links content' response to transmission in a specific news context [5]. Video research additionally shows that social referrals, creator following, text, shot scale, and growth trajectories can affect measured outcomes [6-9]. Douyin research further shows that popularity indicators are not interchangeable [10]. Without platform data, this study cannot infer engagement or virality from content resemblance.

The visible weaknesses concern meaning across layers. Humor can depend on the timing between setup and reaction. Informational content can fail when one label is wrong. Demonstrations can become misleading when a step or object changes. Storytelling can preserve an emotional relationship while altering causality. Multimodal analysis treats visual objects, written language, sound, and events as separate yet interacting units [1]. A coherent frame can therefore coexist with an inaccurate label or an altered sequence.

5.3. Limitations

Several limitations define the conclusions. Source links, account identities, publication dates, view histories, shares, and selection records were unavailable. Viral status could not be verified. The model names, complete prompts, one-attempt rule, shared defaults, and absence of post-production were documented, while the precise model builds, generation dates, seeds, and default parameter values were not. M4 combined visual and extensive textual guidance, and all prompt packages differed in information quantity. One reviewer used standardized frames without a validated scale. Audio was not transcribed or scored. Generated clips were also shorter and more uniform than many references. Future work should archive source URLs and performance histories, record every parameter and seed, use repeated generations, and vary one prompt component at a time. Several independent raters should evaluate full clips and audio with a defined codebook. Audience testing and platform data would be required for claims about engagement or virality.

6. Conclusion

This exploratory comparison indicates that the two-model pipeline can reproduce the broad topic, genre, focal subject, and emotional direction of many human reference clips. Correspondence was stronger for stable situations with one clear relationship. Fidelity declined when a reference required multi-step action, factual text, comic timing, or a longer causal sequence. The four prompt packages produced overlapping strengths and weaknesses. The multimodal reference package appeared closer to source composition in several cases, yet it did not consistently preserve action or event order.

The results answer the research questions within a limited evidence boundary. Doubao-Seedance-2.0 reproduced high-level characteristics more reliably than detailed procedures and narratives after Qwen3.8-Max converted references into prompts. Differences appeared across packages, although the one-shot design and unequal information prevent a causal ranking. This study systematically documents the reference fidelity of generated videos within the constructed corpus but does not evaluate content virality nor identify a universal performance advantage for vivid textual descriptions, keywords, or visual prompts.

References

- [1] Grzenkiewicz, M., & Wildfeuer, J. (2025). Addressing TikTok's multimodal complexity: A multi-level annotation scheme for the audio-visual design of short video content. *Digital Scholarship in the Humanities*, 40(4), 1143-1166. <https://doi.org/10.1093/llc/fqaf047>
- [2] Lubart, T. (2005). How can computers be partners in the creative process: Classification and commentary on the special issue. *International Journal of Human-Computer Studies*, 63(4-5), 365-369. <https://doi.org/10.1016/j.ijhcs.2005.04.002>
- [3] Rezwana, J., & Maher, M. L. (2023). Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction*, 30(5), Article 67, 1-28. <https://doi.org/10.1145/3519026>

- [4] Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try and fail to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Article 437, pp. 1-21). <https://doi.org/10.1145/3544548.3581388>
- [5] Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2), 192-205. <https://doi.org/10.1509/jmr.10.0353>
- [6] Broxton, T., Interian, Y., Vaver, J., & Wattenhofer, M. (2013). Catching a viral video. *Journal of Intelligent Information Systems*, 40(2), 241-259. <https://doi.org/10.1007/s10844-011-0191-2>
- [7] France, S. L., Vaghefi, M. S., & Zhao, H. (2016). Characterizing viral videos: Methodology and applications. *Electronic Commerce Research and Applications*, 19, 19-32. <https://doi.org/10.1016/j.elerap.2016.07.002>
- [8] Jiang, L., Miao, Y., Yang, Y., Lan, Z., & Hauptmann, A. G. (2014). Viral video style: A closer look at viral videos on YouTube. In *Proceedings of the International Conference on Multimedia Retrieval* (Glasgow, United Kingdom, pp. 193-200). <https://doi.org/10.1145/2578726.2578754>
- [9] Ling, C., Blackburn, J., De Cristofaro, E., & Stringhini, G. (2022). Slapping cats, bopping heads, and Oreo shakes: Understanding indicators of virality in TikTok short videos. In *Proceedings of the 14th ACM Web Science Conference* (pp. 164-173). <https://doi.org/10.1145/3501247.3531551>
- [10] Chen, Z., He, Q., Mao, Z., Chung, H.-M., & Maharjan, S. (2019). A study on the characteristics of Douyin short videos and implications for edge caching. In *Proceedings of the ACM Turing Celebration Conference China* (pp. 1-6). <https://doi.org/10.1145/3321408.3323082>