JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 2, pp. 347-356, Feb. 2026

Check for updates

AIGC 장시간 시퀀스 영상 일관성 실증 및 비교연구

장 예 한[1] · 선 심 이[2] · 정 진 헌[3*]
[1]동국대학교 영상대학원 석사과정
[2]중국 제남대학교 미술디자인학원 교수, 동국대학교 영상대학원 박사
[3]동국대학교 영상대학원 교수

Empirical and Comparative Study of Long-Sequence Video Consistency in AIGC

Yuhan Zhang[1] · Xinyi Shan[2] · Jean-Hun Chung[3*]

[1]Master's Course, Multimedia Department, Graduate School of Digital Image and Contents, Dongguk University, Seoul 04620, Korea
[2]Lecturer, School of Fine Arts and Design, University of Jinan, Shandong 250022, China
[3]Professor, Multimedia Department, Graduate School of Digital Image and Contents, Dongguk University, Seoul 04620, Korea

[요 약]

생성형 AI의 부상과 함께 AI 기반 영상 합성 기술은 영화, 광고 및 뉴미디어 분야에서 혁신적인 도구로 자리 잡았다. 그러나 복잡한 장면 생성 시 시간적 불연속성, 물리적 일관성 부족, 스타일 변형과 같은 문제는 여전히 해결해야 할 과제로 남아 있다. 본 연구는 지맹(JiMeng), 비두(Vidu), 컬링(Keling) AI를 대상으로 숲/동물, 도시/거리, 실내/인물, 해변/자연, 공상과학/도시, 제품/전시 등 6가지 시나리오에서 비교 분석을 수행하였다. 통일된 프롬프트와 표준화된 프레임 연속성 전략을 적용하여 기본 설정에서 5초 길이의 영상(16:9)을 생성하였다. 연구 결과, 지맹은 도시, 공상과학, 제품 장면에서, 컬링은 자연 환경에서, 비두는 실내 인물의 표정 묘사에서 각각 우수한 성능을 보였다. 본 연구는 플랫폼 평가 패러다임을 제안하고 시나리오별 강점을 도출함으로써 AI 주도 미디어 환경의 창의적 활용을 위한 핵심적인 기술 지침을 제공한다.

[Abstract]

With the rise of generative AI, AI-based video synthesis has emerged as a transformative tool in film, advertising, and new media. However, complex scenes continue to face challenges, such as temporal discontinuity, lack of physical consistency, and style shifts. This study conducts a comparative analysis of JiMeng, Vidu, and Keling AI across six scenarios: forest/animals, city/street, indoor/people, beach/nature, sci-fi/city, and product/exhibition. Using unified prompts and a standardized frame-continuity strategy, 5-s videos (16:9) were generated under default settings. Results show that JiMeng performs best in urban, sci-fi, and product scenes; Keling excels in natural environments; and Vidu stands out in indoor character expressions. This study proposes a platform evaluation paradigm and highlights scenario-specific strengths, providing essential technical guidance for creative applications in the AI-driven media landscape.

색인어 : AI 영상 생성, 내용 연속성, 물리적 일관성, 스타일 안정성, 연속 영상
Keyword : AI Video Generation, Content Coherence, Physical Consistency, Style Stability, Sequential Video

http://dx.doi.org/10.9728/dcs.2026.27.2.347

Received 31 October 2025; **Revised** 02 January 2026
Accepted 29 January 2026

***Corresponding Author; Jean-Hun Chung**
Tel: ※ 개인정보 표시제한
E-mail: evengates@gmail.com

Ⅰ. Introduction

With the rapid advancement of artificial intelligence, image-to-video technology has become widely used as a key tool in film, advertising, education, and other fields. Yet, current platforms face challenges in long-form generation, including limited content coherence, motion inconsistency, style instability, and varying usability. This study systematically compares three Chinese platforms—JiMeng AI, Vidu AI, and Keling AI—by generating multi-stage continuous videos using identical initial images and last-frame continuation. Evaluations cover content coherence, physical consistency, style stability, and efficiency. The findings provide an objective benchmark of platform capabilities, offering references for long-video generation and future research.

Ⅱ. Theoretical Background and Platform Overview

2-1 AI Video Generation and Image-to-Video Mechanism

With the rapid development of generative AI, video generation has become a key direction in digital content production. Current approaches include Text-to-Video (T2V), Image-to-Video (I2V), and Reference-to-Video (R2V). Among these, I2V offers unique advantages for visual creation and long-video research, as it maps spatial information from static images into temporal dynamics to enable continuous scene expansion. However, existing I2V technologies face challenges such as limited content coherence, insufficient physical consistency, and unstable style across long or multi-segment generations. This study examines three representative platforms to compare their performance in last-frame continuation for long-video generation.

2-2 Platform Overview

1) JiMeng AI

Developed by Shenzhen Lianmeng Technology and promoted by ByteDance, JiMeng AI (Dreamina) launched in 2024 to “turn inspiration into finished content instantly.” It supports AI image and video generation with features like story mode, first-to-last frame continuity, lip-sync, and camera control. Its Seaweed and OmniHuman models enhance dynamic human portraits, while Action Imitation allows fine-grained control, serving creators with efficient workflows and a rich material library.

2) Vidu AI

Vidu AI, by Beijing Shengshu Technology and Tsinghua University, is China’s first long-duration, high-consistency, high-dynamics video model, globally launched in July 2024. Using the U-ViT architecture combining Diffusion and Transformer, it supports Reference-, Image-, and Text-to-Video generation up to 8 seconds at 1080P. The 1.5 update added multi-image reference for multi-subject consistency, improving stability in complex scenes.

3) Keling AI

Developed by Kuaishou AI, Keling AI supports videos up to 1080P, 2 minutes (30fps), and flexible aspect ratios. Since 2024, it introduced Image-to-Video, web access, and multimodal upgrades (Kling 2.0, Ketu 2.0). Optimized for long, high-resolution, dynamic scenes, it is stable and widely used alongside JiMeng AI and Vidu AI.

2-3 Related Work and Evaluation Frameworks

Recent studies in AI video generation have introduced new paradigms for evaluating image-to-video (I2V) and text-to-video (T2V) systems.

Recent evaluation frameworks for generative video models have focused on perceptual alignment, motion quality, and physical plausibility.

VMBench proposes a perception-aligned benchmark for assessing motion realism and temporal coherence [4], while VideoPhy specifically targets physical commonsense consistency in generated videos[3].

In addition, VideoGen-Eval introduces an agent-based evaluation system to enhance robustness and interpretability in video generation assessment[2].

Fan et al. proposed the AIGCBench benchmark, defining eleven metrics across four dimensions: control-video alignment, motion effectiveness, temporal consistency, and perceptual quality.

Zhang et al. developed the DropletVideo-10M dataset, focusing on dynamic camera motion and

causal VAE-based modeling to improve long-term temporal coherence.

Other works such as Phantom emphasize text-image-video triplet alignment to enhance subject consistency and cross-modal fidelity[6].

Compared to these, the present study contributes a "last-frame continuation" strategy for video extension and a cross-platform, multi-scenario comparative framework[7].

This design enables a controlled evaluation of three commercial systems under identical conditions, providing reproducible insights into temporal stability, stylistic preservation, and platform adaptability—an approach rarely documented in prior research[5].

Ⅲ. Research Methodology and Experimental Design

3-1 AI Video Generation and Image-to-Video Mechanism

In this study, a set of core variables was systematically defined and controlled to ensure comparability and stability across different platforms in image-to-video generation. Given the probabilistic nature of generative diffusion models, particular attention was paid to distinguishing between controllable inputs and platform-dependent internal parameters.

From the perspective of experimental control, prompt text, initial images, video duration, and output resolution were treated as constants. All prompts were fixed in structure, descriptive granularity, and stylistic intent across platforms. Initial images were standardized in terms of content composition and resolution, and all generated video clips were constrained to a fixed duration of 5 seconds with a uniform aspect ratio, thereby eliminating variations caused by temporal length or spatial scale.

In contrast, platform-specific generation mechanisms constituted the primary comparative variables. These include differences in model architecture, training data distribution, and internal sampling strategies. Importantly, two critical parameters in generative diffusion-based systems—random seed values and guidance scale (classifier-free guidance strength)—were not manually controlled in this study but instead left at each platform's default settings.

While seed control is commonly used to ensure deterministic reproducibility within a single model, cross-platform comparisons face inherent limitations, as most commercial platforms do not expose seed values or guidance scale parameters to users. As a result, this study deliberately adopted default configurations (JiMeng Video 3.0, Kling 2.1, Vidu Q1) to reflect real-world usage conditions and to evaluate platform behavior as experienced by general creators. Consequently, variations in generated results are understood to partially arise from differences in each platform's internal default guidance strength, sampling temperature, and stochastic initialization strategies.

Rather than treating this variability as experimental noise, the study considers it an intrinsic characteristic of platform performance, revealing how each system balances semantic alignment, motion stability, and visual diversity under identical external inputs. This approach allows for a realistic assessment of platform-level generation tendencies while maintaining controlled input conditions.

1) Constants and Variables

Controlled variables include prompt text (fixed style, length, and semantic structure), initial images (standardized resolution and comparable visual composition), and video duration. Specifically, all input images were normalized to the same resolution, and all generated videos were fixed at 5 seconds in length, ensuring that resolution and temporal scale did not influence the comparative results.

Comparative variables primarily consist of the video generation platform itself. Although camera motion was specified textually within prompts, internal motion synthesis and temporal interpolation were handled independently by each platform's model architecture and sampling process. Due to the lack of user-level access to seed numbers and guidance scale parameters, these factors were implicitly treated as platform-dependent variables rather than externally controlled ones.

Under this framework, the experiment focuses on evaluating how different platforms respond to identical prompts and initial frames, thereby highlighting

variations in coherence, motion plausibility, and stylistic stability that emerge from distinct model designs and default parameter settings.

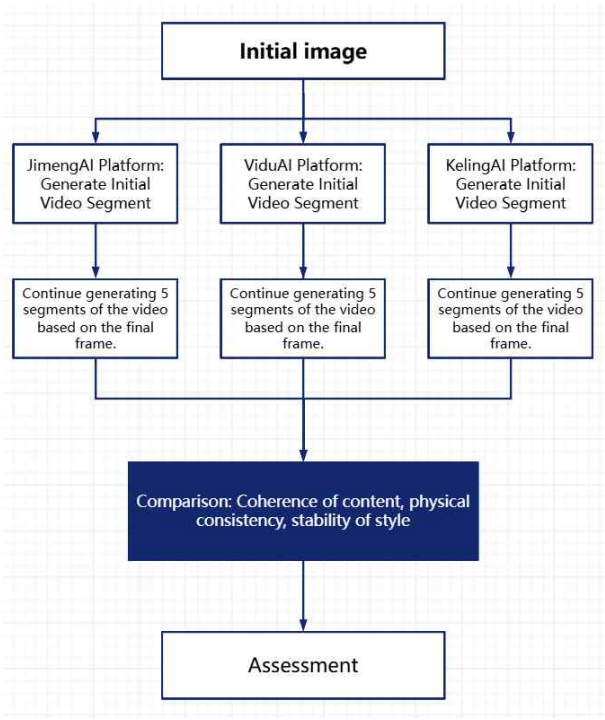


Fig. 1. Flowchart

2) Prompt and Initial Image Design

Prompt design adhered to principles of clarity, specificity, and cinematic structure, explicitly describing scene context, subject attributes, motion behavior, camera movement, and visual style. For example, the prompt "A red fox walks through a misty pine forest, tracked by a smooth camera; lighting is soft, style realistic" integrates environmental description, subject action, and camera dynamics within a single coherent instruction.

Each experimental scenario employed a tailored prompt paired with an initial keyframe image, which served as the visual anchor for subsequent generation. This design ensured alignment between intended semantics and generated output while supporting consistent comparison across platforms during multi-stage video continuation.

3-2 Tail-Frame Continuation Strategy

This study employs a tail-frame-based approach to generate long videos. The target video is divided into consecutive 5-second segments, generated sequentially. For the first segment, a clip is created using the initial frame (from prompt or input image) and the textual prompt. The last frame of each segment is then extracted and used as the starting input for the next, until all segments are concatenated into a complete video.

Tail frames are extracted via standard tools (e.g., OpenCV) and serve as reference images for continuation. While motion and content within each segment are guided by prompts, platform-specific randomness makes this strategy an effective test of consistency. This method evaluates each platform's ability to maintain coherence across segments and highlights issues such as content drift, repetition, or quality degradation during long-sequence generation.

3-3 Experimental Procedure and Configurations

Six representative scenarios were selected, each described by six sequential prompts (see Appendix). For each prompt, a 5-second video clip (16:9) was generated under default platform settings. Three major Chinese AI video generation platforms were chosen: JiMeng (Video 3.0), K e ling (Video 2.1), and Vidu (Q1), each representing distinct model architectures and optimization mechanisms. The latest public versions were adopted, and all experiments were conducted under identical hardware conditions to eliminate performance bias.

Following the tail-frame continuation strategy, the last frame of each segment was extracted and used as the initial input for the subsequent prompt. Iteratively, six clips were generated and concatenated to form a continuous video, ensuring temporal coherence and consistency of subjects and backgrounds across segments.

3-4 Platform Architecture Analysis

The three AI video generation platforms—JiMeng AI, Keling AI, and Vidu AI—differ fundamentally in their technical architectures and generation mechanisms, leading to distinct performance characteristics across scenarios.

JiMeng AI employs ByteDance's Seedream model, built upon a Diffusion Transformer (DiT) architecture that integrates multimodal alignment through

cross-modal rotational positional encoding. This design enables high adaptability to text-image prompts and nuanced control of camera motion.

Keling AI (version 2.0) follows a Sora-like DiT structure, replacing conventional U-Net with a Transformer backbone and introducing a redesigned VAE for enhanced spatiotemporal coherence and physical realism.

In contrast, Vidu AI adopts the novel U-ViT framework that fuses Diffusion and Transformer mechanisms, achieving end-to-end one-shot long video generation. Its model excels in representing realistic physics, lighting, and character expression continuity.

Moreover, Keling AI supports multi-subject and multi-frame consistency control, JiMeng AI features "motion painting board" tools for camera trajectory editing, and Vidu AI focuses on character identity and stylistic stability.

As a result, JiMeng AI tends to perform better in narrative flow and cinematic composition, Keling AI demonstrates superior motion dynamics and realism, while Vidu AI achieves stable temporal coherence and expressive control of human subjects.

Ⅳ. Results and Analysis

4-1 Overall Performance of the Three Platforms

This study compared Keling, JiMeng, and Vidu across six scenarios using questionnaires and subjective ratings. Results show that Keling performed best in natural scenes such as forests and coastlines, producing clear subjects and high physical realism. JiMeng excelled in urban, sci-fi, and product demonstration scenarios, with smooth transitions and coherent narratives. Vidu showed strengths in indoor human-centered scenes, generating natural expressions and consistent styles. Overall, each platform demonstrated domain-specific advantages across different scene categories.

4-2 Comparison of Content Coherence

Content coherence was evaluated in terms of narrative continuity and temporal consistency. Keling achieved stable performance in natural scenes (e.g., forest, coastline), maintaining visible subjects and natural transitions. JiMeng outperformed others in urban scenes, ensuring smooth trajectories and seamless camera shifts, and also excelled in sci-fi and product demonstrations. Vidu showed the highest coherence in indoor human scenarios, with natural action-to-expression transitions. In summary, Keling is more effective for natural scene narratives, JiMeng maintains coherence in complex urban contexts, while Vidu performs slightly better in human-related continuity.

4-3 Comparison of Physical Consistency

Physical consistency measures how well dynamic elements follow natural laws. Keling excelled in realistic scenes: forest leaves and pine needles moved naturally under gravity and wind, and seaside waves and reflections matched prompts, outperforming JiMeng and Vidu. Urban vehicle motion and simple indoor/product scenes were handled similarly by all platforms. These results align with prior studies (Gu et al., PhyWorldBench; Bansal et al., VideoPhy), confirming Keling's advantage in simulating physical details like water flow and gravity.

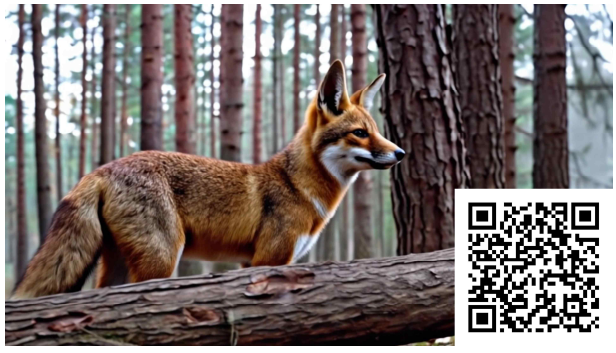


Fig. 2. Video frames generated by Kling AI in the forest - animal scenario

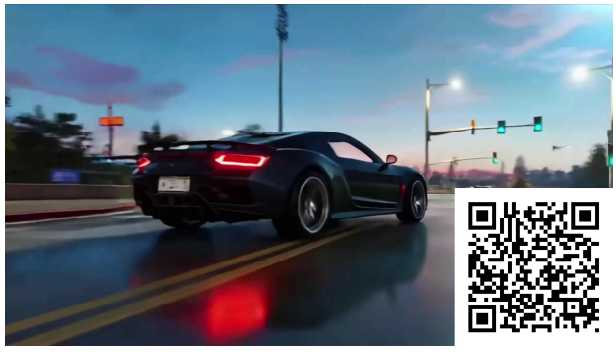


Fig. 3. Video frames generated by JiMeng AI in the city - street scene

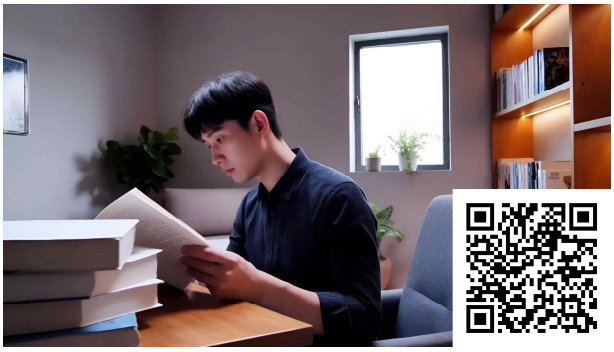


Fig. 4. Video frames generated by Vidu AI in the indoor - human scene

Fig. 5. Video frames generated by Keling AI in the seaside·natural scene

Fig. 6. Video frames generated by Jimeng AI in the Sci-fi·urban scene

Fig. 7. Video frames generated by Jimeng AI in the product·showcase scene

4-4 Comparison of Style Stability

Style stability assesses whether videos consistently reflect the prompt's visual style. In forest scenes, Keling slightly outperformed Vidu in lighting and tonal consistency. In urban scenes, JiMeng maintained the cyberpunk neon style best, while Vidu and Keling leaned toward realism. Indoor scenes favored Vidu's warm, natural lighting, whereas others showed occasional inconsistencies. Seaside scenes saw Keling produce the most coherent sunset and beach tones. In sci-fi scenes, JiMeng maintained stable cool-blue neon lighting; Keling and Vidu deviated slightly. For product showcases, all platforms used a minimalist aesthetic, with JiMeng offering the most refined metallic details. Overall, platforms captured stylistic intent, though frame-to-frame consistency remains a challenge.

4-5 Camera Motion Performance

Camera motion assesses whether videos follow the prompt's specified movements. In forest scenes, only Keling achieved smooth, accurate tracking; JiMeng and Vidu showed occasional jitter. Urban lateral and overhead car shots were best handled by JiMeng, while Keling sometimes lagged. Indoor pans and push-ins were smoothest in Vidu, with others showing abrupt jumps. Seaside left-to-right pans were effective in Keling and JiMeng, but Vidu was slightly stiff. Sci-fi upward pulls and dive-follow shots were best in JiMeng; other platforms had minor deviations. Product rotations were well handled by Keling and JiMeng, with JiMeng more stable in close-ups. Overall, JiMeng excelled in urban camera control, Keling in natural scenes, aligning with Li et al.'s "camera" prompt design principle.

V. Results and Discussion

5-1 Interpretation of Results

Under identical experimental conditions—including the same initial images, unified prompts, fixed clip duration (5 seconds, 16:9), last-frame continuation strategy, and default model configurations (JiMeng Video 3.0, Kling 2.1, and Vidu Q1)—the three platforms

exhibited distinct generation behaviors across the six scenarios. The observed inconsistencies in object interaction and motion plausibility are consistent with prior findings on physical commonsense and perception-aligned motion evaluation[3],[4].

The evaluation was conducted using four criteria: content coherence, physical consistency, style stability, and prompt-image alignment, based on frame-level visual analysis and aggregated subjective survey scores.

Fig. 8. Frame comparison

Fig. 9. Frame comparison

As illustrated in Fig. 8 and Fig. 9, noticeable differences emerged in object continuity, motion smoothness, and camera behavior when comparing representative frames from each platform. Some systems preserved object identity and spatial relationships more effectively across successive segments, while others exhibited visible drift in texture, lighting, or motion trajectories after multiple continuations.

These variations can be primarily attributed to differences in model architecture, training data distribution, and sampling strategy. JiMeng AI and Kling AI are both based on DiT-style diffusion architectures; however, Kling 2.1 incorporates enhanced multimodal text-image alignment and an improved VAE module, contributing to more stable object deformation and smoother physical transitions—particularly in scenes involving natural motion such as water, foliage, or atmospheric effects.

This tendency is further supported by prior studies emphasizing that large-scale video models trained with strong physical priors and real-world motion distributions perform better in natural dynamics modeling, including fluid motion, soft-body deformation, and environmental interactions[3]. In addition, Kling's development documentation and technical releases highlight optimization for long-duration, high-resolution, real-world dynamic scenes, which aligns with its superior performance in forest and coastal scenarios.

In contrast, Vidu AI adopts a U-ViT architecture that performs long-range spatiotemporal inference within a unified Transformer framework. This design emphasizes global temporal coherence, enabling more consistent preservation of human facial features, posture, and identity across multiple continuation segments. Such characteristics correspond with prior findings that Transformer-based temporal modeling is advantageous for maintaining subject identity and facial consistency in human-centered video generation [1],[2].

Training data bias further explains stylistic tendencies observed in the results. JiMeng AI demonstrated stronger performance in urban, sci-fi, and product showcase scenes, which can be plausibly linked to its training emphasis on cinematic composition, camera control, and semantically dense visual content commonly found in urban, commercial, and science-fiction media. Industry reports and platform descriptions indicate that JiMeng prioritizes narrative-driven generation and prompt controllability, particularly for Chinese-language inputs and structured scenes involving architecture, vehicles, and designed objects.

Such characteristics are consistent with prior observations that models trained on highly structured visual domains—such as urban environments and synthetic or semi-synthetic sci-fi imagery—exhibit

higher prompt fidelity and compositional stability in these contexts[1]. Consequently, JiMeng's stronger semantic alignment and camera control capabilities are most evident in city, sci-fi, and product scenarios.

It should also be noted that diffusion-based generation inherently involves stochastic sampling. Even under identical prompts and default settings, repeated inference can yield non-identical outputs. This probabilistic nature explains minor variations observed within the same platform and underscores the broader technical challenge of achieving deterministic consistency in long-sequence AI video generation.

Across the six evaluated scenarios, the comparative results can be summarized as follows: JiMeng AI achieved the highest overall coherence and prompt fidelity in city, sci-fi, and product scenes; Kling AI produced the most physically convincing results in natural environments such as forests and coastlines; and Vidu AI most effectively maintained character identity and facial expression consistency in indoor human-centered scenes.

Because all experiments were conducted using default platform versions, these findings are time-sensitive. Model updates or parameter tuning may alter performance; however, under controlled conditions, the observed differences remain consistent and reproducible within the scope of this study.

5-2 Implications for Creative Applications

The results provide several practical implications for AI-assisted video creation. First, platform selection can be optimized according to scene characteristics: Kling AI is better suited for generating physically grounded backgrounds and natural environments, such as forests, coastlines, or atmospheric scenes; Vidu AI excels in character-driven segments requiring stable facial expressions and identity preservation; and JiMeng AI performs strongly in structured storytelling, camera-controlled sequences, and product-oriented visuals.

Building on these findings, the study proposes a hybrid, multi-platform production workflow that goes beyond simple performance comparison. For example, environmental backgrounds can be generated using Kling AI to ensure realistic physical motion (e.g., water, foliage, lighting), while foreground human characters are generated separately using Vidu AI to maintain facial and identity consistency. These components can then be composited in post-production, with JiMeng AI optionally used for transitional shots, narrative continuity, or camera-guided sequences. Such a pipeline enables creators to systematically exploit complementary platform strengths, rather than relying on a single model for all stages of production.

Furthermore, the results confirm that prompt precision and keyframe continuity significantly influence short-term coherence. However, even with carefully designed prompts and last-frame continuation, long-sequence generation still exhibits gradual semantic and visual drift. This limitation indicates that manual intervention, keyframe correction, or post-editing remains necessary for professional-grade outputs.

Finally, the rapid iteration and frequent version updates of commercial platforms highlight the importance of systematic version control, parameter documentation, and reproducibility practices in both creative and research contexts. For model developers, the findings suggest a need for improved temporal stabilization, more transparent controllable parameters, and enhanced long-range consistency mechanisms to support reliable long-form video generation.

5-3 Limitations

Despite its contributions, this study has several limitations. First, the experimental scope was limited to six scenarios and a fixed set of prompts, which constrains the generalizability of the conclusions. Second, evaluation relied primarily on aggregated subjective assessments; although this approach reflects real user perception, it introduces potential bias. Future research should integrate objective quantitative metrics such as SSIM, LPIPS, or FVD to strengthen reliability.

Third, all results are dependent on specific model versions and default configurations, making them sensitive to rapid platform updates. Fourth, while camera motion was included in prompt descriptions, it

was not systematically isolated or tested as an independent variable. Additionally, limited repetition reduced the ability to statistically analyze generation randomness and reproducibility. Finally, ethical and copyright considerations related to AI-generated video content were beyond the scope of this study but warrant dedicated investigation in future work.

Ⅵ. Conclusion

This study employed a last-frame continuation strategy to systematically compare three mainstream AI video generation platforms—JiMeng, Vidu, and Kling—across six representative scenarios (forest, city, indoor, seaside, sci-fi, product display). Using unified prompts and fixed-length clips (5s × 6 segments), we conducted a quantitative evaluation through online surveys and expert ratings, focusing on content coherence, physical consistency, style stability, and generation efficiency. Results indicate that JiMeng excels in urban/sci-fi and product display scenes for prompt fidelity and narrative coherence; Kling demonstrates clear advantages in simulating natural physical details (e.g., water splashes, wind, falling leaves); and Vidu performs best in maintaining character expressions and detail consistency. Methodologically, this work contributes a reproducible last-frame continuation paradigm and multi-dimensional evaluation framework, offering empirical guidance for platform selection and benchmarking. Nonetheless, limitations include restricted scene and prompt coverage, reliance on subjective scoring, and time sensitivity to model versions. Future research should expand sample size and repetitions, incorporate objective metrics (SSIM, LPIPS, FVD), systematically examine camera motion as an independent variable, and explore multi-platform collaborative workflows as well as ethical and copyright compliance frameworks, thereby enhancing the stability and applicability of long-sequence AI video generation.

Reference

[1] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, ... and Z. Liu, “VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models,” arXiv:2411.13503, 2025. https://doi.org/10.48550/arXiv.2411.13503

[2] Y. Yang, K. Fan, S. Sun, H. Li, A. Zeng, F. Han, ... and Z.-J. Zha, “VideoGen-Eval: Agent-based System for Video Generation Evaluation,” arXiv:2503.23452, 2025. https://arxiv.org/html/2503.23452v1

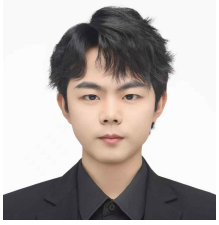
[3] H. Bansal, Z. Lin, T. Xie, Z. Zong, M. Yarom, Y. Bitton, ... and A. Grover, “VideoPhy: Evaluating Physical Commonsense for Video Generation,” arXiv:2406.03520, 2025. https://doi.org/10.48550/arXiv.2406.03520

[4] X. Ling, C. Zhu, M. Wu, H. Li, X. Feng, C. Yang, ... and X. Chu, “VMBench: A Benchmark for Perception-Aligned Video Motion Generation,” arXiv:2503.10076, 2025. https://doi.org/10.48550/arXiv.2503.10076

[5] S. Ma, H. Xu, M. Li, W. Geng, Y. Wang and M. Wang, “POS: A Prompts Optimization Suite for Augmenting Text-to-Video Generation,” arXiv:2311.00949, 2023. https://doi.org/10.48550/arXiv.2311.00949

[6] S.-X. Zhang, H. Wang, D. Huang, X. Li, X. Zhu and X.-C. Yin, “VCapsBench: A Large-scale Fine-grained Benchmark for Video Caption Quality Evaluation,” arXiv:2505.23484, 2025. https://doi.org/10.48550/arXiv.2505.23484

[7] H. Han, S. Li, J. Chen, Y. Yuan, Y. Wu, C. T. Leong, ...and Y. Ni, “Video-Bench: Human Aligned Video Generation Benchmark,” arXiv:2504.04907, 2025. https://doi.org/10.48550/arXiv.2504.04907



장예한(Yuhan Zhang)

2024.06 : Department of Digital Media Art Design, Qilu University of Technology (BFA)
2025.03 : Department of Multimedia, Graduate School of Digital Image and Contents, Dongguk University (MFA)

2020.08～2024.06: Student, School of Art and Design, Qilu University of Technology, China
2025.03～Present: Graduate Student, School of Fine Arts and Design, Dongguk University, Seoul, South Korea
※Research Interests : Contents Design, 3D Computer Graphic, Intelligent Product Development, AI Art, Interaction Design, etc

선심이(Xinyi Shan)

2014.02 : Department of Video Design, Pyeongtaek University (BFA)
2016.02 : Department of Multimedia, Graduate School of Digital Image and Contents, Dongguk University (MFA)
2023.08 : Department of Multimedia, Graduate School of Digital Image and Contents, Dongguk University (Ph.D Degree)

2014.10～2016.04: ABITS Communications
2016.12～2018.07: ableMEDIA
2018.08～2022.02: Associate Professor, School of Art, Shandong Yingcai University, China
2024.02～Present: Lecturer, School of Fine Arts and Design, University of Jinan, Shandong, China
※Research Interests : Contents Design, 3D Computer Graphic, Intelligent Product Development, AI Art, Interaction Design, etc

정진헌(Jean-Hun Chung)

1992년 : Department of Visual Design, College of Fine Arts, Hongik University KOR (BFA)
1999년 : Computer Arts, Academy of Art University USA (MFA)

2001년～Present: Professor of Multimedia Department, Graduate School of Digital Image and Contents, Dongguk University
※Research Interests : VR, Contents Design, 3D Computer Graphic, Computer Animation, Visual Effects, AI Art, etc