한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 31, No. 5, P. 17~27

Infinite Video Generation with Cinematic Camera Trajectory Control

Jungmin Lee[1†] Jongwon Choi[1] Jaewon Song[2*]

[1]Dept. of Advanced Imaging, GSAIM, Chung-Ang University, Republic of Korea
[2]DEXTER STUDIOS, Republic of Korea
[†]Work done during internship at DEXTER STUDIOS

시네마틱 카메라 제어 기반 무제한 길이 비디오 생성

이정민[1†] 최종원[1] 송재원[2*]

[1]중앙대학교 첨단영상대학원 영상학과
[2] (주)덱스터 스튜디오
[†]본 연구는 (주)덱스터 스튜디오 인턴십 기간 중 수행되었음

jngmnlee@vilab.cau.ac.kr, choijw@cau.ac.kr, jaewon.song@dexterstudios.com

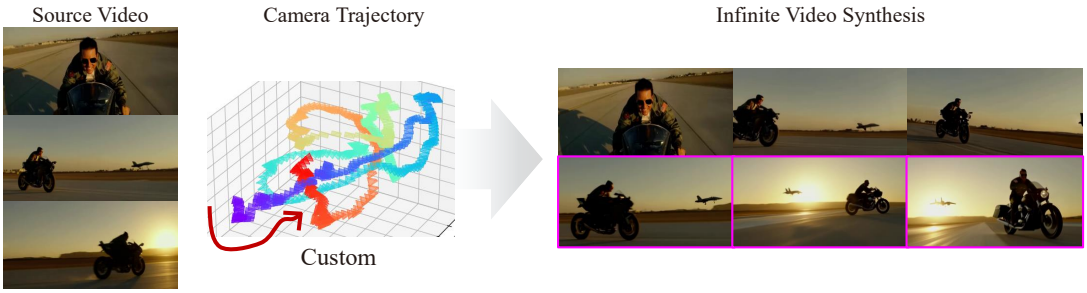


Figure 1: **Examples synthesized by our framework.** We regenerate the source video using custom trajectories with unlimited length. The pink-highlighted section shows extended frames not in the source video.

Abstract

Current video generation faces critical limitations in camera control and sequence length, which restrict its applicability in filmmaking. Existing approaches support only basic camera movements within short 5-second sequences. We present an infinite video generation framework that enables the creation of unlimited-length sequences with precise camera trajectory control. We provide control over position, rotation, and movement intensity for arbitrary cameras, along with 12 standard camera presets for ease of use. Experimental results demonstrate robust applications of diverse camera trajectories across various content types. Our work connects conventional cinematography with synthetic video production, expanding possibilities for filmmaking applications.

요 약

현재 비디오 생성 기술은 카메라 제어의 제한성과 짧은 시퀀스 길이로 인해 영화 제작에 실질적으로 활용되기 어렵다. 기존 방법들은 5초 이내의 짧은 영상에서 단순한 카메라 움직임만 구현하여 복잡한 영상 서사 구성에 한계가 있다. 본 연구는 정밀한 카메라 궤적 제어를 통해 길이 제한 없이 연속적인 시퀀스를 생성하는 장편 비디오 프레임워크를 제안한다. 시스템은 3차원 위치, 회전각, 움직임 강도를 정밀 조절하며 12가지 표준 카메라 프리셋을 제공한다. 실험 결과 다양한 콘텐츠에서 카메라 궤적이 안정적으로 적용됨을 확인하였다. 본 연구는 전통 촬영 기법과 합성 비디오 기술을 연결하여 영화 제작 분야의 AI 비디오 생성 활용 가능성을 확장한다.

Keywords: Video Generation, Camera Control, Film Production, Deep Learning, Computer Vision
키워드: 비디오 생성, 카메라 제어, 영화 제작, 답러닝, 컴퓨터 비전

*corresponding author: Jaewon Song/DEXTER STUDIOS, Republic of Korea (jaewon.song@dexterstudios.com)

Received : 2025.07.30./ Review completed : 1st 2025.10.02./ Accepted : 2025.10.14.
DOI : 10.15701/kcgs.2025.31.5.17
ISSN : 1975-7883(Print)/2383-529X(Online)

1 Introduction

Video generation has demonstrated remarkable progress in recent years. Breakthroughs in diffusion models [1] have enabled photorealistic video synthesis with exceptional temporal consistency. Google's released Veo 3 [2] demonstrates impressive capabilities in generating high-quality video content. These technological advances have opened new possibilities for cinematic content creation, with synthetic video showing significant potential in film production workflows.

However, the current video generation encounters significant limitations in camera control. Users often struggle to implement specific camera trajectories or movements with the precision required for cinematic storytelling. Camera motion typically occurs arbitrarily during the generation process. More problematically, modifying the camera trajectory of already-generated videos remains impossible. This limitation forces creators to rely on trial-and-error approaches that require repeatedly generating videos until satisfactory results emerge, a process that is both time-consuming and resource-intensive.

Camera-controllable video generation models have emerged to address these fundamental issues. However, these approaches still suffer from serious constraints that limit practical utility in professional contexts. First, precise camera control remains impossible. The functionality remains limited to operating within a few presets only. Second, the length of generatable videos remains limited to 5 seconds. The limitation makes the models unsuitable for creating long-take footage required in films. The limitation creates barriers for creators seeking to implement complex visual expressions.

To solve the problems, we propose a framework that simultaneously supports infinite-length video generation and customizable camera movements. Our framework enables filmmakers to create cinematic content using standardized camera presets and custom trajectories tailored to specific creative requirements, as illustrated in Fig. 1. Our approach provides the following key contributions:

- We propose an infinite video generation framework that enables the creation of extended cinematic sequences and long-take shots.
- We develop a customizable trajectory system with precise control over camera path and speed derived from film production.
- We introduce camera presets that provide standard camera movements for filmmakers.
- We demonstrate the practical applicability of our framework across diverse content types and cinematographic scenarios.

2 Related Works

2.1 Video Generative Models

Video generation has experienced rapid advancement in recent years. The initial approaches combined 2D spatial processing with 1D temporal modeling [3, 4, 5], then evolved into 3D attention mechanisms [6, 7]. Key architectural advances include the transition from U-Net [8] to DiT [9] and MMDiT [10] architectures. Furthermore, optimization strategies have evolved from DDPM approaches [1, 11] to flow matching methodologies [12, 13].

Variational auto-encoders (VAEs) have improved along with text encoding capabilities [14, 15, 16]. Advanced vision-language models, such as BLIP [17] and LLaVA [18], have significantly improved automated video descriptions and enabled more accurate text-to-video generation. Despite these improvements in video fidelity, current systems still face significant challenges for commercial deployment, primarily due to limited user control over camera viewpoints and scene composition.

Recent commercial video generation systems have demonstrated impressive capabilities in automated content creation [19, 20, 21]. However, these systems primarily focus on text-driven generation with limited user control over specific cinematic aspects such as camera movements and scene composition. While some commercial tools offer basic camera presets, they provide only approximations of cinematic movements without fine-grained control over trajectory parameters such as movement intensity, speed variation, or precise spatial positioning. The gap between automated generation capabilities and precise creative control remains a significant challenge for professional filmmaking, where directors require exact camera trajectories.

2.2 Camera-Controlled Video Generation

Camera-controlled video generation has emerged as a critical research area following the success of text-to-video models [7, 22, 23]. Early methods achieve basic camera control through high-level instructions. MovieGen [24] uses text descriptions to control camera motion, while MCDiff [25] and DragNUWA [26] enable control through user-provided strokes. AnimateDiff [4] introduces motion LoRAs [27] to learn camera movement patterns from augmented datasets.

Recent approaches focus on control using camera parameters as conditional input. MotionCtrl [28] directly injects 6 DoF camera extrinsic into diffusion models through fine-tuning on video-camera pair datasets. CameraCtrl [29] employs specialized encoding to represent camera origin and ray directions with improved accuracy. CVD [30] proposes cross-video synchronization modules,

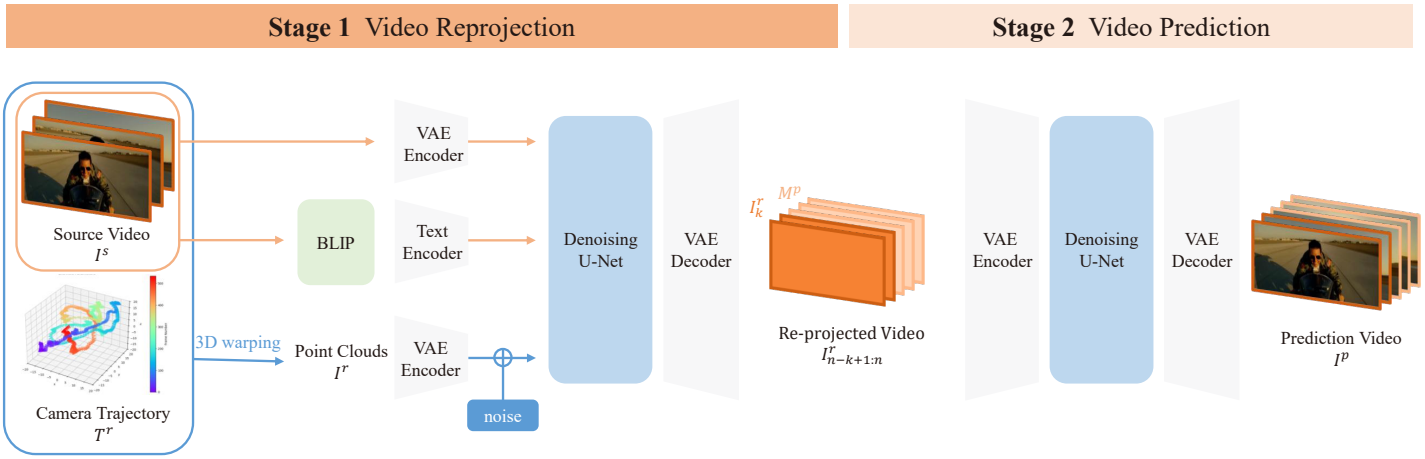


Figure 2: **Overview of framework.** Our framework operates through two alternating stages: Stage 1 performs video reprojection to achieve precise camera trajectory control, while Stage 2 conducts video prediction for temporal extension by using the re-projected frames as conditioning input to generate subsequent frames. This dual-stage approach enables the generation of infinite-length video sequences with accurate camera control throughout the entire sequence.

and AC3D [31] investigates camera motion knowledge within diffusion transformers.

Recent camera-controllable re-generation enables capturing dynamic scenes from source videos with specified camera trajectories. ReCamMaster [32] uses token concatenation for camera-controlled scene reproduction, while TrajectoryCrafter [33] employs a dual-stream diffusion model that combines point cloud renders and source videos. However, existing methods face critical limitations as they depend on user-provided strokes or predefined camera presets, preventing accurate trajectory re-rendering for specific user requirements. Additionally, generated videos remain limited to short durations, making them unsuitable for long-take applications required in professional film production.

3 Preliminaries

3.1 Video Diffusion Models

Video diffusion models [5, 6] operate in the latent space using a VAE to encode video frames into lower-dimensional representations. Given a video sequence $I = \{I_i\}_{i=1}^n \in \mathbb{R}^{n\times 3\times h\times w}$, a pre-trained VAE $\mathcal{E}$ maps frames to latent representations $z_0 = \mathcal{E}(I)$. The diffusion process corrupts the latent representation by adding Gaussian noise:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t$ controls the noise schedule and t denotes the diffusion timestep.

3.2 Masked Video Diffusion

Masked video diffusion enables selective conditioning on visible frames while generating masked regions. Given a binary mask sequence $M = \{M_i\}_{i=1}^n$ that indicates frame visibility, we compute the masked latent code as:

$$\tilde{z}_0 = z_0 \odot M, \tag{2}$$

where $\odot$ denotes element-wise multiplication. During diffusion, the denoising network reconstructs content in masked regions while preserving visible frame information. This approach generates content in occluded and future regions to enable infinite video generation with precise camera control.

4 Method

Our proposed method enables precise trajectory control over long-form video generation through a recurrent process. Our approach consists of two main components: video reprojection for novel view synthesis and video prediction for future frames. We adopt TrajectoryCrafter [33] as the base model for video reprojection and Seine [34] as the base model for video prediction. Fig. 2 shows an overview of our framework.

4.1 Video Reprojection

The video reprojection module converts source videos to novel viewpoints through depth-driven geometric rendering and masked diffusion. Given a source video $I^s = \{I_i^s\}_{i=1}^n$, we initial calculate

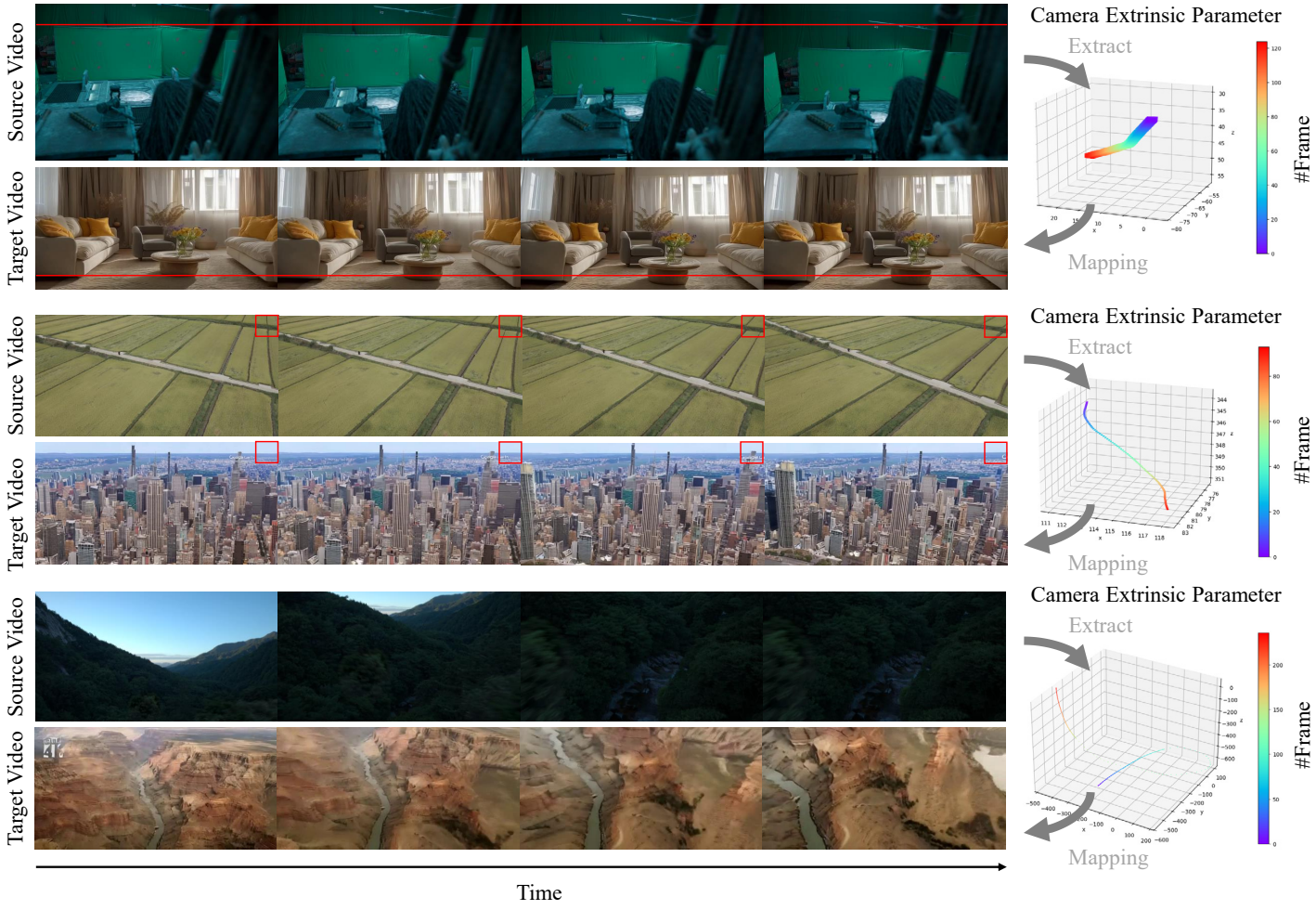


Figure 3: **Videos regenerated with real film production trajectories.** We demonstrate trajectory mapping by extracting camera extrinsic parameters from actual film productions and applying these cinematic trajectories to diverse content types, including synthetic scenes, indoor environments, aerial footage, and natural landscapes. Each row shows the source video, target video, and the corresponding 3D trajectory plot that illustrates camera positions and orientations throughout the sequence, with color-coded temporal progression. The red boxes highlight key frames demonstrating precise camera movement control. Please zoom in for more details.

depth maps $D^s = \{D_i^s\}_{i=1}^n \in \mathbb{R}^{n\times h\times w}$ [35]. We then construct a point cloud sequence $P = \{P_i\}_{i=1}^n$ by applying inverse perspective projection:

$$P_i = \Phi^{-1}([I_i^s, D_i^s], K), \tag{3}$$

where Φ^{-1} represents the inverse perspective projection, and $K \in \mathbb{R}^{3\times3}$ denotes the camera intrinsic matrix. Using this point cloud, we render novel views $I^r = \{I_i^r\}_{i=1}^n$ following a specified camera trajectory $T^r = \{T_i^r\}_{i=1}^n \in \mathbb{R}^{n\times4\times4}$:

$$I_i^r = \Phi(T_i^r \cdot P_i, K), \tag{4}$$

where Φ performs the perspective projection. Since point cloud reconstruction occurs within the coordinate frame of the source camera, T_i^r represents the transformation matrix relative to the initial viewpoint. The projected rendered images I^r contain visible gaps due to occlusion effects and boundary limitations, which the module identifies through mask sequences $M^r = \{M_i^r\}_{i=1}^n$.

The camera trajectory $\mathbf{T}^r$ can be specified through two approaches: user-provided c2w matrices from external camera tracking systems, or predefined camera movement presets. For preset-based generation, we implement 12 standard cinematographic movements through parameterized transformations of the initial camera pose. Rotational presets (pan, tilt, arc) apply Euler angle rotations with angular range $\theta_{max} = \alpha \cdot s$, where α is the base rotation angle specific to each movement type and s is the intensity parameter. Translational presets (dolly, translation) apply linear displacement with magnitude $d_{max} = \beta r \cdot s$, where r represents the scene radius estimated from depth maps and β is a preset-specific scaling factor. Zoom presets operate by modifying the camera intrinsic matrix $\mathbf{K}$ through focal length interpolation while maintaining fixed spatial position. Each trajectory frame is

computed as $T_i^r = \mathbf{R}(\theta_i)\mathbf{T}(d_i) \cdot T_0^r$, where $\mathbf{R}$ and $\mathbf{T}$ denote rotation and translation matrices with parameters linearly interpolated across frames.

Although s technically represents movement intensity or range rather than temporal speed, it effectively controls the visual perception of camera velocity. The actual perceived speed results from the combination of s and the total number of frames n, where identical intensity values produce faster motion with fewer frames and slower motion with more frames. Low intensity values produce subtle, cinematic motions suitable for dramatic scenes, while high values create dynamic, energetic camera work for action sequences. We apply the intensity scaling uniformly across all trajectory points to maintain geometric consistency while achieving the desired visual dynamics.

The reprojection module utilizes I^r and M^r as conditioning inputs to guide video synthesis following the camera trajectory. The projected views I^r and the occlusion masks M^r provide spatial constraints, while the source sequence I^s provides appearance details via a reference-conditioning diffusion architecture. This architecture incorporates cross-attention layers that connect the projected content with the source material, thereby preserving visual consistency throughout the viewpoint transformation.

4.2 Video Prediction

The video prediction module extends the reprojected sequence temporally using masked diffusion. Taking the final k frames $I^r_{n-k+1:n}$ from the reprojected sequence, we form an N-frame input sequence where the first k frames provide conditioning and the remaining $N - k$ frames are predicted by the module. The input sequence undergoes encoding using a pre-trained VAE to obtain latent representations. Subsequently, we selectively condition on visible frames while predicting masked regions. To enable conditional generation, we apply binary masks $M^p = \{M_i^p\}_{i=1}^N$ where conditioning frames are visible ($M_i^p = 1$ for $i = 1, ..., k$) and target frames are masked ($M_i^p = 0$ for $i = k{+}1, ..., N$). The diffusion model then generates the masked future frames while conditioning on the visible context from the reprojection stage.

Once the video prediction module generates the extended sequence $I^p = \{I_i^p\}_{i=k+1}^N$, this predicted output serves as the new input source video for the subsequent reprojection stage. The framework then processes the predicted frames I^p through the video reprojection pipeline described in Section 4.1. This recurrent alternation between temporal extension and spatial reprojection enables our framework to generate long-form videos while maintaining precise control over the camera trajectory throughout the entire sequence.

The recurrent process terminates based on a predetermined frame count specified by the user. Given a target video length of F frames and the reprojection module output of n frames per cycle, our framework calculates the required number of iterations as $\lceil \frac{F-n}{N-k} \rceil$, where $N - k$ represents the number of newly predicted frames per cycle. This deterministic termination ensures precise control over the final video length while maintaining computational efficiency and predictability.

5 Experiments

5.1 Experimental Settings

Dataset. For camera trajectory data, we utilized trajectories captured during real film and drama productions. We obtained trajectories using filming and tracking equipment that provides precise camera motion parameters, including position, rotation, and intrinsic parameters. We collect distinct camera trajectories covering various motion patterns such as drone shots, crane movements, and handheld camera work from our original footage of *Along with the Gods*, *The Haunted Palace*, and *Head over Heels*.

For source video data, we constructed a diverse test dataset consisting of 30 video clips from three different sources: (1) 10 movie clips, (2) 10 real-world videos, and (3) 10 synthetic videos. The movie clips include scenes from our original footage and trailers from *Top Gun* and *Toy Story*. The real-world videos include sequences rendered from Google Earth Studio [36], and clips captured from *Around The World 4K* [37]. The synthetic videos are generated by Luma AI [21].

Implementation. We implemented our method using PyTorch [38] and performed all experiments on a single NVIDIA A100 GPU. We adopt TrajectoryCrafter [33] and Seine [34] as the base models. Our method does not require additional training beyond these pretrained models. During inference, the framework requires at least 28GB of VRAM to process 49-frame sequences. In our recurrent pipeline, we set the number of conditioning frames to $k = 5$. For camera trajectory visualization in our experiments, we utilize CameraCtrl [29] to render the camera paths and movements.

5.2 Results

Fig. 3 demonstrates our trajectory mapping method using camera movements from real film production. We apply these trajectories to three different target videos: synthetic room scenes (Luma AI), aerial urban scenes (Google Earth Studio), and natural landscapes (Around The World 4K). The 3D plots show camera po-

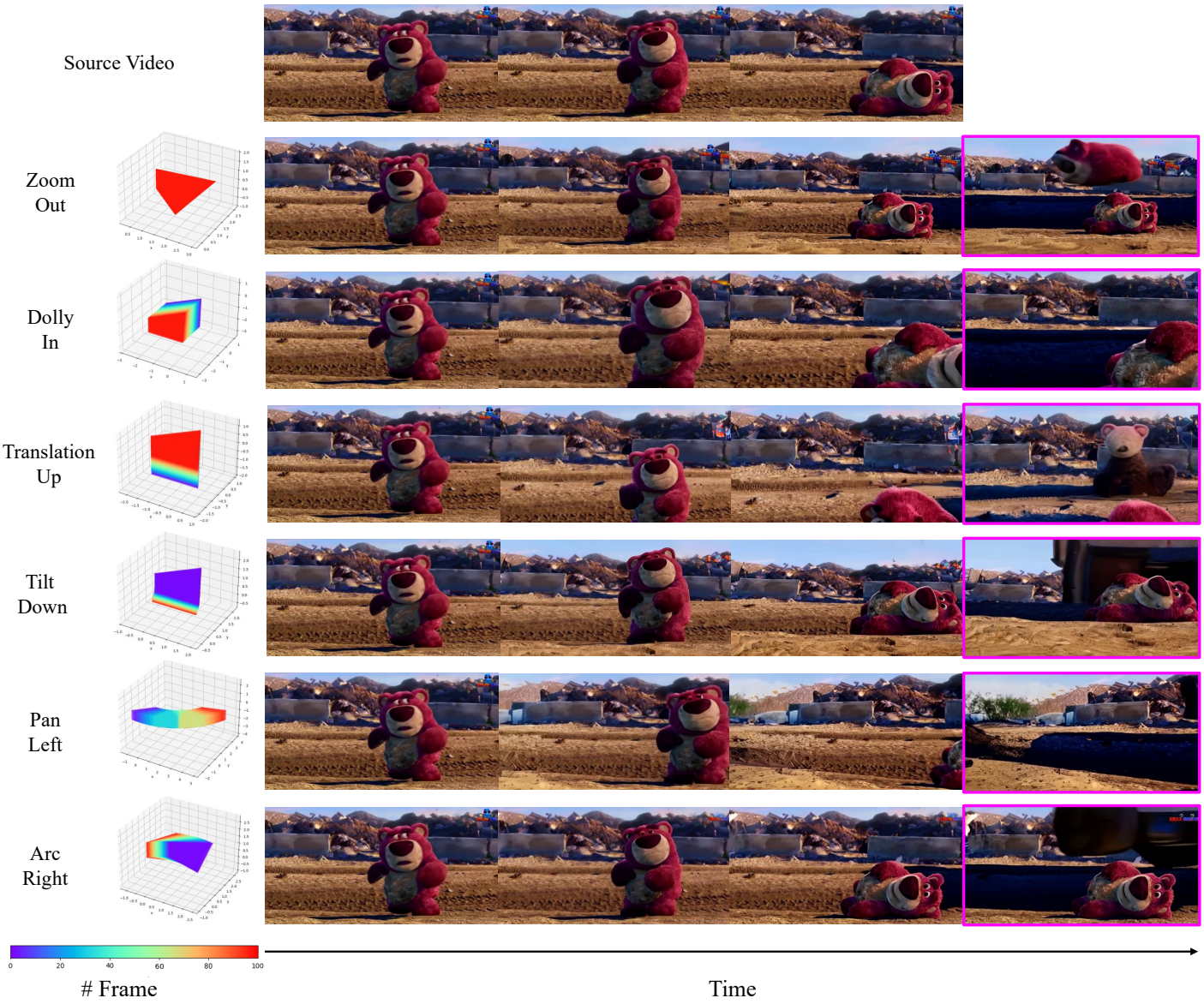


Figure 4: **Videos regenerated with camera presets.** We visualize 6 out of 12 available camera presets to demonstrate the diversity of cinematographic movements. Each row shows a different camera preset applied to the same source video content. The 3D trajectory plots on the left illustrate the spatial movement patterns for each preset, with color-coded temporal progression. The pink-highlighted sections show frames generated using camera presets via the prediction module, demonstrating smooth transitions and consistent camera control throughout the extended sequence.

sitions throughout each trajectory, with color gradients indicating temporal progression and cones marking camera orientation. Our approach excels at capturing and reproducing subtle cinematic motions that are important in professional filmmaking. The synthetic room scenes and aerial urban scenes particularly demonstrate our ability to handle fine-grained camera adjustments, such as gentle drift, micro-movements, and gradual transitions. Our method successfully transfers both pronounced and nuanced camera dynamics across diverse content types while preserving the original motion timing. This capability enables more sophisticated cinematic expression compared to previous approaches that primarily emphasize visually striking camera motions.

Fig. 4 presents the camera preset-based video generation results. We generate videos using predefined camera presets that correspond to standard cinematic movements. The results demonstrate how each preset creates distinct visual effects while maintaining content consistency. The pink-highlighted sections indicate frames generated by the prediction module using camera presets. This approach enables filmmakers to achieve camera work without requiring complex manual trajectory specifications. Fig. 5 illustrates the effect of speed parameters on pan movement execution. We apply three speed settings (0.1, 1.0 and 10) to identical pan trajecto-

Figure 5: **Visual comparison of pan mode at different intensities.** We demonstrates pan movement with varying intensity parameters (s=0.1 for subtle motion, s=1 for standard motion, and s=10 for rapid movement) to illustrate adjustable camera movement control. Each row shows the same pan trajectory applied with different speed settings.

Table 1: **Comparison of video regeneration capabilities.** The table compares three key capabilities: maximum generatable video length in frames, number of available camera movement presets, and support for dynamic intensity control of camera motions. Our method is shown in blue.

Method	# Frames	# Camera Presets	Intensity Control
GCD [39]	14	0	
TrajectoryCrafter [33]	49	3	
ReCamMaster [32]	81	10	
Ours	∞	**12**	✓

ries, effectively showing how speed control enables fine-tuning of movement dynamics while preserving spatial accuracy and trajectory fidelity. The results highlight the system's ability to generate both cinematic slow and rapid camera movements from the same preset configuration.

Tab. 1 demonstrates the capabilities of our method in video regeneration. For the number of frames, our approach supports unlimited renderable frames, far exceeding GCD [39] (14 frames), TrajectoryCrafter [33] (49 frames), and ReCamMaster [32] (81 frames). For the number of camera presets, our method offers 12 presets (e.g., zoom in/out, dolly in/out, translation up/down, tilt up/down, pan left/right, and arc left/right), outperforming GCD (0 presets), TrajectoryCrafter (3 presets), and ReCamMaster (10 presets). For speed control, our method is the only one that supports dynamic adjustment of camera movement speed, allowing users to accelerate or decelerate camera motions as needed.

To quantitatively evaluate temporal consistency, we measure frame-to-frame coherence using CLIP similarity [41] and LPIPS [42] metrics. We apply identical source videos and camera trajectories across all methods. As shown in Tab. 2, our method achieves superior temporal consistency with CLIP similarity of 0.9575 and LPIPS of 0.0409. Our method builds upon TrajectoryCrafter for video reprojection and extends the framework to infinite-length generation through recurrent prediction.

Table 2: **Quantitative evaluation of temporal consistency.** We report CLIP Similarity (higher is better) and LPIPS (lower is better) for frame-to-frame consistency in generated videos.

Method	CLIP Similarity ↑	LPIPS ↓
GCD [39]	0.7831	0.2810
ViewCrafter [40]	0.8919	0.1209
ReCamMaster [32]	0.9341	0.0631
Ours	**0.9575**	**0.0409**

6 Discussion

While our method successfully generates long-form videos with precise camera trajectory control, we observe certain limitations as video duration increases. As the generation process extends through multiple recurrent cycles, the visual content gradually diverges from the source video due to accumulated variations in each reprojection-prediction cycle. This divergence becomes more pronounced with longer durations, as the model progressively relies less on the initial visual context and more heavily on text prompt guidance. In extended sequences, the text prompt becomes the primary semantic anchor, while the source video's influence diminishes. Consequently, while camera trajectories remain accurately controlled, the visual content may evolve beyond the original scene's characteristics, resulting in semantic drift from the initial visual reference.

To address this limitation, future work could implement a 3D scene system that accumulates point clouds from previous reprojection stages. Currently, our video prediction module only conditions the most recent k frames, leading to gradual content drift from the source video. A more robust approach would maintain a 3D scene representation by storing and integrating point clouds generated throughout the entire sequence. When predicting future frames, the system would leverage this accumulated 3D knowledge to identify existing scene geometry and only generate point clouds for newly visible or occluded regions. This approach would construct a progressive 3D scene reconstruction as the camera moves, enabling more consistent 2D rendering that preserves the original scene's spatial and semantic characteristics. Particularly in loop closure scenarios where the camera returns to its starting point, accumulated geometric drift may cause spatial inconsistencies between the initial and final frames, requiring global optimization mechanisms to ensure 3D coherence across the entire trajectory. Such a scene-aware framework would maintain stronger visual

coherence with the source content while still enabling unlimited-length generation with precise camera control.

7 Conclusion

We present a video generation framework that addresses two fundamental challenges in current video synthesis, specifically limited sequence length and insufficient camera control. Our method generates unlimited-length videos with precise trajectory control by leveraging real film camera movements through a recurrent alternation between video reprojection and temporal prediction. The system provides 12 standard camera presets that cover professional cinematographic movements with dynamic intensity control, demonstrating robust performance across diverse content domains, from synthetic scenes to real-world footage. This approach successfully bridges traditional filmmaking techniques with AI-driven video synthesis, enabling authentic reproduction of cinematic motion while maintaining temporal consistency.

This study opens significant possibilities for the film industry and content creation more broadly. Our framework enables cost-effective film previsualization and rapid prototyping of cinematic sequences, potentially democratizing high-quality video production tools. By establishing a connection between professional cinematography and synthetic video generation, we expand creative possibilities and lay the groundwork for professional-grade AI video tools. While challenges such as semantic drift in extended sequences present opportunities for future research, our approach represents a crucial step toward sophisticated camera-controlled video synthesis systems that can meet the evolving demands of modern content creation workflows.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00229451, Interoperable Digital Human (Avatar) Interlocking Technology Between Heterogeneous Platforms)

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] DeepMind, "Veo 3," 2025.

[3] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, *et al.*, "Make-a-video: Text-to-video generation without text-video data," *International Conference on Learning Representations*, 2023.

[4] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *International Conference on Learning Representations*, 2024.

[5] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[6] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[7] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[9] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[10] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[12] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *International Conference on Learning Representations*, 2022.

[13] W. Jin, Q. Dai, C. Luo, S.-H. Baek, and S. Cho, "Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

[14] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You, "Open-sora: Democratizing efficient video production for all," *arXiv preprint arXiv:2412.20404*, 2024.

[15] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang, *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[16] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[17] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[19] Runway, "Gen-4: Next generation video generation," 2024, accessed: 2025-07-10. [Online]. Available: https://runwayml.com/gen-4

[20] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and C. Zhang, "Video generation models as world simulators," 2024, technical Report. [Online]. Available: https://openai.com/research/video-generation-models-as-world-simulators

[21] LumaAI, "Dream machine," 2025.

[22] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7310–7320.

[23] W. Menapace, A. Siarohin, I. Skorokhodov, E. Deyneka, T.-S. Chen, A. Kag, Y. Fang, A. Stoliar, E. Ricci, J. Ren, *et al.*, "Snap video: Scaled spatiotemporal transformers for text-to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7038–7048.

[24] A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang, *et al.*, "Movie gen: A cast of media foundation models," *arXiv preprint arXiv:2410.13720*, 2024.

[25] T.-S. Chen, C. H. Lin, H.-Y. Tseng, T.-Y. Lin, and M.-H. Yang, "Motion-conditioned diffusion model for controllable video synthesis," *arXiv preprint arXiv:2304.14404*, 2023.

[26] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan, "Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory," *arXiv preprint arXiv:2308.08089*, 2023.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models." *International Conference on Learning Representations*, vol. 1, no. 2, p. 3, 2022.

[28] Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan, "Motionctrl: A unified and flexible motion controller for video generation," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[29] H. He, Y. Xu, Y. Guo, G. Wetzstein, B. Dai, H. Li, and C. Yang, "Cameractrl: Enabling camera control for video diffusion models," in *The Thirteenth International Conference on Learning Representations*, 2025.

[30] Z. Kuang, S. Cai, H. He, Y. Xu, H. Li, L. J. Guibas, and G. Wetzstein, "Collaborative video diffusion: Consistent multi-video generation with camera control," *Advances in Neural Information Processing Systems*, vol. 37, pp. 16 240–16 271, 2024.

[31] S. Bahmani, I. Skorokhodov, G. Qian, A. Siarohin, W. Menapace, A. Tagliasacchi, D. B. Lindell, and S. Tulyakov, "Ac3d: Analyzing and improving 3d camera control in video diffusion transformers," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 875–22 889.

[32] J. Bai, M. Xia, X. Fu, X. Wang, L. Mu, J. Cao, Z. Liu, H. Hu, X. Bai, P. Wan, *et al.*, "Recammaster: Camera-controlled generative rendering from a single video," *International Conference on Computer Vision*, 2025.

[33] M. YU, W. Hu, J. Xing, and Y. Shan, "Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models," *International Conference on Computer Vision*, 2025.

[34] X. Chen, Y. Wang, L. Zhang, S. Zhuang, X. Ma, J. Yu, Y. Wang, D. Lin, Y. Qiao, and Z. Liu, "Seine: Short-to-long video diffusion model for generative transition and prediction," in *The Twelfth International Conference on Learning Representations*, 2023.

[35] W. Hu, X. Gao, X. Li, S. Zhao, X. Cun, Y. Zhang, L. Quan, and Y. Shan, "Depthcrafter: Generating consistent long depth sequences for open-world videos," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2005–2015.

[36] Google, "Google earth studio," https://www.google.com/earth/studio/, 2018, accessed: 2025-07-07.

[37] ATWFilms, "Grand canyon 4k," YouTube https://www.youtube.com/watch?v=huqJUghX26Y, 2014, accessed: 2025-07-07.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[39] B. Van Hoorick, R. Wu, E. Ozguroglu, K. Sargent, R. Liu, P. Tokmakov, A. Dave, C. Zheng, and C. Vondrick, "Generative camera dolly: Extreme monocular dynamic novel view synthesis," in *European Conference on Computer Vision*. Springer, 2024, pp. 313–331.

[40] W. Yu, J. Xing, L. Yuan, W. Hu, X. Li, Z. Huang, X. Gao, T.-T. Wong, Y. Shan, and Y. Tian, "Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis," *arXiv preprint arXiv:2409.02048*, 2024.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[42] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[43] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *ACM Transactions on Graphics*, pp. 1214–1220, 2006.

[44] P. Alliez, D. Cohen-Steiner, O. Devillers, B. Levy, and M. Desbrun, "Anisotropic polygonal remeshing," *ACM Transactions on Graphics*, pp. 485–493, 2003.

[45] X. Shi, Z. Huang, F.-Y. Wang, W. Bian, D. Li, Y. Zhang, M. Zhang, K. C. Cheung, S. See, H. Qin, *et al.*, "Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[46] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

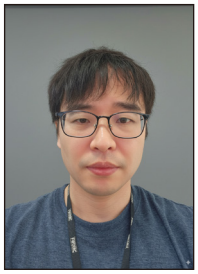
〈저자소개〉

이정민

- 2020 국립공주대학교 문화재보존과학과 이학사
- 2023 한국전통문화대학교 문화유산전문대학원 공학석사
- 2024.03~현재 중앙대학교 첨단영상대학원 박사과정
- 관심분야: 콘텐츠 AI, 디지털 헤리티지, 컴퓨터비전
- https://orcid.org/0009-0009-1172-7209

최종원

- 2012 KAIST 전기전자공학과 공학학사
- 2014 KAIST 전기전자공학과 공학석사
- 2018 서울대학교 전기정보공학부 공학박사
- 2018.06~2020.03 삼성 SDS, AI Research Center, Research Scientist
- 2020.03~2024.02 중앙대학교 첨단영상대학원 조교수
- 2024.03~현재 중앙대학교 첨단영상대학원 부교수
- 관심분야: 콘텐츠 AI, 컴퓨터비전
- https://orcid.org/0000-0001-9753-8760



송재원

- 2011 KAIST 문화기술대학원 공학석사
- 2017 KAIST 문화기술대학원 공학박사
- 전 ㈜디지털아이디어 R&D 연구소장
- 전 ㈜엔진비주얼웨이브 R&D 연구소장
- 현 홍익대학교 영상커뮤니케이션대학원 겸임교수
- 현 ㈜덱스터스튜디오 R&D 연구소장
- 관심분야: 디지털 휴먼, 모션 캡처, 리타겟팅
- https://orcid.org/0009-0004-6081-7618