Appendix

Anonymous ECCV 2026 Submission

Paper ID #4999

A Additional Related Work

A.1 Camera Trajectory Generation

Traditional approaches in automatic cinematography employ predefined rules and optimization techniques to control camera behavior. Early works utilized analytical methods based on script annotations to automate basic movements like pan, tilt, and zoom, ensuring subject visibility within frames [9, 22]. Optimization approaches manage complex camera intrinsics (e.g., focal length [17]) and extrinsics [6, 19], integrating aesthetic principles like composition [20], viewpoint continuity [1], and action coherence [31]. Recent advancements leverage neural networks for enhanced flexibility: reinforcement learning agents optimize camera trajectories via human preference scores [7] or reward functions balancing aesthetics and fidelity [33]. Transformers [28] and GAN architectures [32] further enhance tracking precision and interactive scene adaptability. While reference-based methods [12, 13] and NeRF-based techniques [16, 26] transfer shooting patterns by aligning human kinematics or heatmap guidance, they primarily focus on long-range aerial shots, lacking fine-grained artistic control. Recent works address this by learning cinematic features from film masterpieces, such as extracting pose-based patterns for camera behavior synthesis [13] or introducing keyframe constraints for stylistic consistency [14].

Emerging learning-based methods leverage diffusion models to generate plausible camera trajectories, yet struggle with multi-shot coherence and stylistic diversity [3, 15]. E.T. [3] focuses on autonomous camera movement but neglects critical cinematographic elements like shot scale, angle control, and multi-segment planning. Cine-AI [5] leverages director-specific datasets to generate game cutscenes with style uniformity, combining user-adjustable storyboards and runtime automation in Unity. Pulp Motion [4] extends this by incorporating framing-aware multimodal camera and human motion generation, emphasizing compositional elements for improved immersion. Collectively, these approaches reduce manual adjustments but fail to holistically integrate narrative context with cinematographic principles, often overlooking emotional and spatial dynamics in virtual environments.

A.2 Advanced Controllable Video Generation

Controllable camera motion in video generation is critical for film and media production. Recent advancements have explored methodologies to enable user-directed camera and object motion control [8, 11]. However, these often rely

on 2D-based frameworks, lacking explicit 3D spatial modeling, which leads to perspective inconsistencies and geometric implausibility. Methods like Direct-a-Video attempt to decouple object and camera motion via sparse spatial constraints but suffer from overlapping artifacts due to 2D attention mechanisms [30].

To address 3D-aware control, recent efforts condition generation on simplified 3D camera trajectories, omitting intrinsic parameters such as focal length [10, 18, 27]. Existing frameworks universally neglect cinematographic principles and lack integration with film-script-driven multi-shot planning or established conventions, highlighting a gap between research and professional workflows. More recent efforts, such as GEN3C [21], introduce 3D-informed world-consistent video generation with precise camera control, leveraging NeRF-like representations for enhanced spatial coherence. CineMaster [24] proposes a 3D-aware framework for cinematic text-to-video generation, incorporating controllable elements like camera paths and scene dynamics. MotionCanvas [29] focuses on cinematic shot design with controllable image-to-video generation, enabling fine-grained motion customization. Frame In-N-Out [23] advances unbounded controllable image-to-video generation, supporting extended sequences with dynamic framing. NewtonGen [34] emphasizes physics-consistent and controllable text-to-video generation, ensuring realistic motion adherence. These innovations aim to bridge the divide by unifying narrative intent, trajectory planning, and aesthetic-aware video synthesis, though challenges in multi-shot coherence and real-time applicability persist.

B Dataset

Most existing 3D virtual camera trajectory datasets are reconstructed from real videos [3, 35], making precise cinematographic annotation difficult and limiting representation to basic motion patterns. Distinct from these datasets, our procedural synthesis framework in Unity fundamentally decouples camera motion from look-at constraints. This separation enables the generation of flexible paths that maintain precise geometric relationships with targets, allowing us to systematically cover complex cinematic compositions that are often underrepresented in real-world video collections.

Trajectory parameterization and augmentation. To ensure precise control over cinematic effects, the raw data in Unity is represented as $C^{\text{raw}} = \{t, q, f\} \in \mathbb{R}^{3+4+1}$, where t, q, and f denote translation, quaternion rotation, and focal length, respectively. This representation supports advanced maneuvers like the dolly-zoom, where camera movement must be synchronized with lens adjustments to maintain composition. For instance, given a camera pose with an initial field of view (FoV) f_1, we adjust the target distance using Equation 1 to strictly preserve the subject's framing size during translation:

$$d_2 = d_1 \cdot \frac{\tan(f_1/2)}{\tan(f_2/2)} \tag{1}$$

Table 1: Dataset comparison. We compare the dataset with (i) RealEstate10k, a pure camera trajectory dataset designed for tasks like novel view synthesis; and (ii) three trajectory datasets captioned with detailed description on lens feature. Notice that Movement Types here only count basic motion types, excluding combinations of basic motions (as they can be easily achieved in game engines).

Dataset	Method	Statistics			Annotations		Intrinsics
		#Samples	#Frames	#Movement	Caption	Tags	(FOV)
RealEstate10k	SLAM / BA	79K	11M	-	✗	✗	✗
E.T.	SfM / HMR	115K	11M	7	✓	✗	✗
DataDoP	SLAM	29K	11M	7	✓	✗	✗
CCD	Synthetic	25K	4.5M	10	✓	✗	✗
LenScript (Ours)	Synthetic	120K	21.6M	13	✓	✓	✓

These raw parameters are subsequently serialized into a RealEstate10k-compatible format $\text{vec}(f, K, [R|t])$, ensuring compatibility with standard geometry-aware pipelines while retaining the procedural precision of the synthetic environment.

Automated Semantic Annotation. Leveraging the procedural nature of our generation pipeline, we eliminate the noise and labor associated with manual labeling. Because every trajectory is synthesized following specific compositional rules, ground-truth motion tags (e.g., Pan Left, Zoom In) are inherently linked to the data. To translate these structured attributes into the natural language required for VLM training, we employ *Qwen-Plus* for large-scale batch inference. We utilize carefully engineered prompts to convert these categorical tags into coherent, descriptive captions. This automated process guarantees high-quality textual alignment for all 120K samples, directly supporting the semantic scoring mechanism described in the method section.

Cinematographic Taxonomy. Our annotation framework adopts a hierarchical cinematographic taxonomy spanning five core dimensions that align with professional filmmaking conventions. This structured classification enables precise control over both camera behavior and visual composition:

- **Shot Motion**: *booms up/down, rotates, trucks, pushed in / out, zoom in / out, dolly zoom in / out, pan, tilt, static.*
- **Shot Scale**: *extreme close, close, medium close-up, medium, long, extreme long.*
- **Shot Direction**: *front, back, left, right, left front, right front, left back, right back.*
- **Shot Angle**: *high-angle, eye-level, low angle.*
- **Screen Property**: *up left, up center, up right, middle left, middle center, middle right, bottom left, bottom center, bottom right.*

C Application, Framework and Discussion

Our framework establishes a unified pipeline for virtual cinematography, integrating script parsing, camera planning, and preference optimization within Unity to

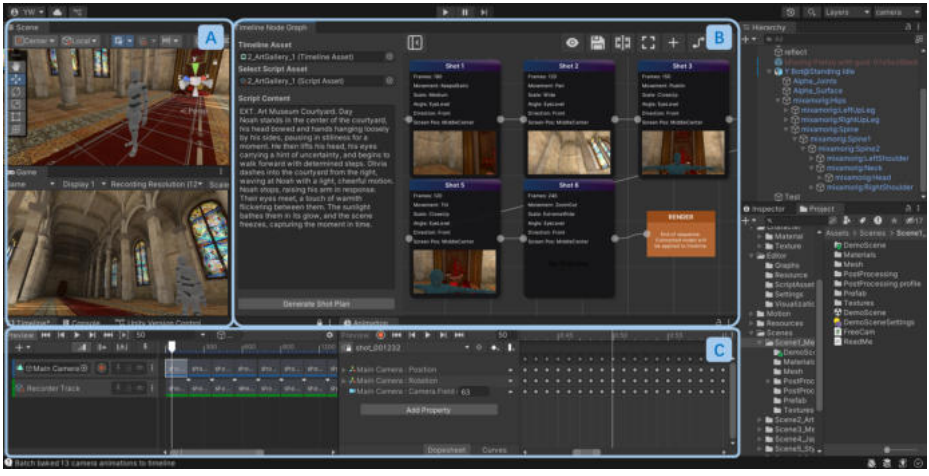


Fig. 1: Unity interface. Panels: (A) 3D Scene and Game views for real-time asset layout and visual preview; (B) node-based storyboard with script editor and shot attribute panels for planning and prompt construction; (C) Timeline and Animation editors for fine-grained trajectory refinement, preview, and batch export supporting VLM scoring and DPO data collection.

enable efficient previsualization and iterative refinement, thereby bridging gaps between trajectory design and downstream video synthesis while accelerating artist workflows.

Script Parsing and Camera Planning. To derive cinematographically valid shot sequences from film scripts and virtual environments, we apply structured schema constraints $\mathbb{T} = \langle l_f, \Theta \rangle$, where $l_f \in \mathbb{R}^+$ denotes frame length, and $\Theta = \{t_i\}_{i=1}^{6}$ specifies cinematographic parameters under typological constraints:

$$t_i \in \bigcup_{j=1}^{m} V_j, \quad V_j = \{v_{j1}, \ldots, v_{jm}\} \tag{2}$$

with V_j as film-theory-curated value sets following the cinematographic taxonomy. User-provided scripts are transformed into structured attributes, which are imported into the game engine and converted to natural-language prompts for guiding trajectory generation, producing pose sequences for video synthesis and preference learning.

Scene-Aware Object Selection. To align virtual environments with narrative contexts, we implement a layer-based selection mechanism that dynamically maps script entities to predefined game engine layers (e.g., characters, objects), ensuring coherent prioritization of shooting targets and scene compositions.

Engine-based Workflow Integration. Our system embeds trajectory generation and preference optimization directly within Unity, establishing an engine-native toolkit that supports seamless integration with industry-standard previsualization workflows. As illustrated in Fig. 1, the Unity interface comprises several interconnected components designed for intuitive artist interaction. (A)

features the 3D scene view and game view, facilitating real-time organization of scene assets, such as props and characters, alongside immediate previews of shooting effects to ensure visual fidelity during planning. (B) introduces a node-based interactive interface, incorporating a film script editor for textual input and refinement, as well as shot attribute editors for specifying parameters like movement, scale, and angle. Users can organize shots into a storyboard via this nodal system, which automatically sequences them into a timeline. Subsequently, leveraging Unity's native timeline and animation editors, artists can perform fine-grained adjustments to virtual camera trajectories enabling precise control over motion dynamics without exiting the environment. We also provide a suite of tools for batch exporting rendered videos from Unity, trajectory previews, and other utilities to facilitate VLM scoring and DPO data construction. This integrated design not only aligns with established production practices used by technical directors and cinematographers but also provides real-time feedback loops, where generated trajectories can be previewed, edited, and exported alongside timeline utilities for Direct Preference Optimization (DPO) data collection, thereby minimizing workflow disruptions and enhancing iterative refinement.

Downstream Applications. By leveraging Unity's rapid rendering capabilities, our framework facilitates efficient previsualization that directly informs downstream video generation pipelines, enabling early validation of camera motions to reduce computational overhead in full synthesis. For instance, Unity-generated preview sequences can be employed in engine-powered diffusion models for cinematic previsualization and video generation [2], allowing artists to refine trajectories before committing to resource-intensive rendering. Furthermore, the optimized camera trajectories serve as explicit conditioning signals during controllable text-to-video or image-to-video generation [10, 21, 25], enhancing geometric consistency and aesthetic quality in outputs from methods, thereby supporting scalable production workflows in virtual cinematography.

D User Study Details

The main paper presents the best-of-4 results for Unity rendering and Wan 2.2 VACE video-to-video transfer. Here, we provide the questionnaire design and the full ranking results.

D.1 Study Design

To assess the perceived cinematic quality of generated trajectories, we collected 34 valid responses, including 11 expert participants with cinematography or video-production experience and 23 general viewers. As shown in Fig. 2, the questionnaire consists of two parts. In the *best-of-4* task, participants view the camera prompt and choose the preferred result among CCD, DIRECTOR, GenDoP, and VERTIGO for five evaluation dimensions. In the *full-ranking* task, participants rank the four methods by overall cinematic quality.

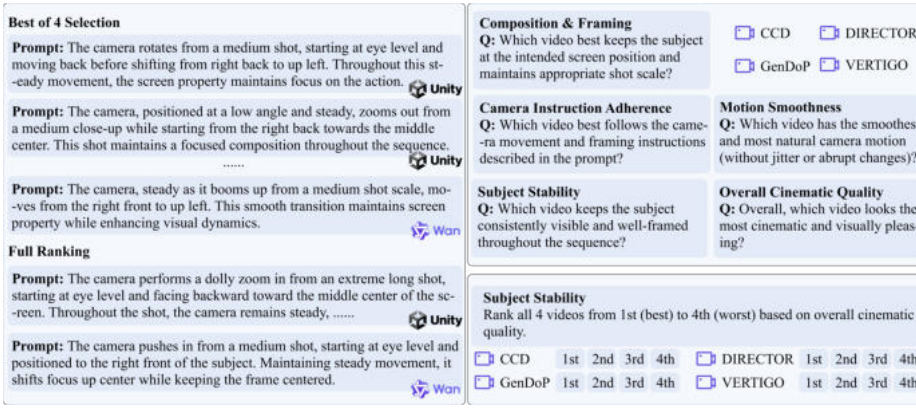


Fig. 2: Questionnaire interface. Top: the best-of-4 interface shows the camera prompt together with five evaluation dimensions, where participants select the best result among four methods for each dimension. Bottom: the full-ranking interface asks participants to rank the four methods by overall cinematic quality.

The study contains five best-of-4 groups and two full-ranking groups. Among them, three best-of-4 groups evaluate Unity-rendered trajectories and two evaluate Wan 2.2 VACE transfer results. For Unity, the five dimensions are composition & framing, motion smoothness, camera instruction adherence, subject stability, and overall cinematic quality. For Wan 2.2 VACE, the five dimensions are composition & framing, temporal consistency, content preservation, transfer quality, and freedom from artifacts.

D.2 Best-of-4 Results

VERTIGO achieves the highest vote share on every evaluated dimension across all five best-of-4 groups. Aggregating the three Unity groups, our method receives 46.1% preference on composition, 50.0% on motion smoothness, 53.9% on instruction adherence, 49.0% on subject stability, and 49.0% on overall cinematic quality, yielding 49.6% overall preference across all Unity questions. Aggregating the two Wan 2.2 VACE groups, VERTIGO obtains 60.3% preference on composition, 55.9% on temporal consistency, 57.4% on content preservation, 55.9% on transfer quality, and 52.9% on artifact suppression, corresponding to 56.5% overall preference. These results are consistent with the main paper and further verify that our preference post-training improves both trajectory-level control and downstream transfer robustness.

D.3 Full-Ranking Results

Table 2 reveals a clear difference between the Unity and transfer ranking tasks. In the Unity ranking group, VERTIGO is the dominant choice, receiving 70.6% first-place votes and the best average rank of 1.47, substantially outperforming GenDoP (2.26), CCD (2.97), and DIRECTOR (3.27).

Method	Unity ranking					Video Gen ranking				
	Avg.↓	1st	2nd	3rd	4th	Avg.↓	1st	2nd	3rd	4th
CCD	2.97	8.8	14.7	47.1	29.4	3.03	5.9	20.6	38.2	35.3
DIRECTOR	3.27	6.1	12.1	30.3	51.5	3.21	8.8	11.8	29.4	50.0
GenDoP	2.26	14.7	55.9	17.6	11.8	2.18	29.4	35.3	23.5	11.8
VERTIGO	1.47	70.6	17.6	5.9	5.9	1.59	55.9	32.4	8.8	2.9

Table 2: Full-ranking results. Percentages indicate how often each method is assigned each rank. A lower average rank is better.

In the Wan 2.2 VACE ranking group, GenDoP achieves the strongest overall ranking performance, receiving 55.9% first-place votes and the best average rank of 1.59, while VERTIGO ranks second with 29.4% first-place votes and an average rank of 2.18. Together with the best-of-4 results, this suggests that dimension-wise judgments and holistic ranking capture complementary aspects of transfer quality: VERTIGO is consistently preferred on targeted criteria such as composition, temporal stability, and artifact suppression, while the single holistic ranking example in VG7 is more favorable to GenDoP.

E Additional Qualitative Results

We present additional qualitative examples across both trajectory-level and video-level comparisons. As illustrated in Fig. 3, VERTIGO generates smooth and cinematographically coherent camera motions that better preserve framing and compositional intent across diverse prompts and scenes. These results further demonstrate the robustness of our post-trained model for trajectory generation in Unity-based previsualization.

Figure 4 further compares all four methods in both Unity rendering and downstream video generation. The improved framing behavior learned by VERTIGO is already visible in the Unity-rendered previews, where the subject remains better positioned and more consistently retained throughout the shot. This advantage transfers to video generation results, yielding more reliable composition, stronger temporal stability, and fewer framing failures than CCD, DIRECTOR, and GenDoP. Together, these examples support our key claim that lightweight render-in-the-loop preference optimization improves not only trajectory quality itself, but also the downstream visual quality of controllable video generation.

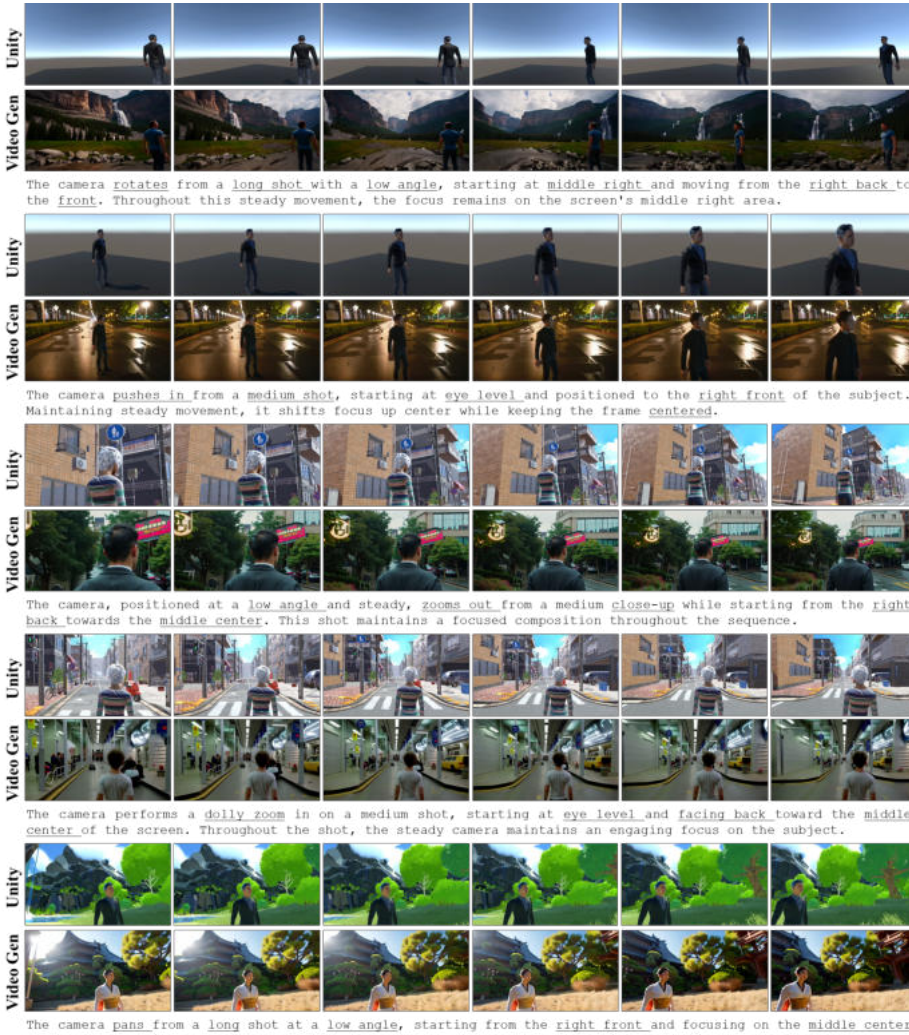
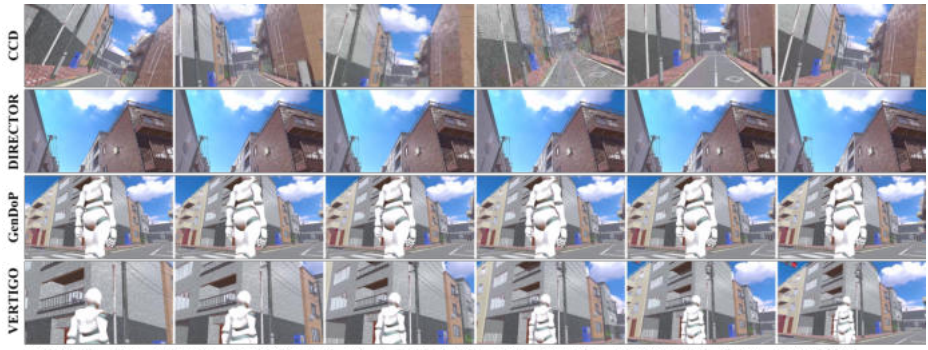


Fig. 3: Additional qualitative comparison of camera trajectories. Across diverse scenes and prompts, VERTIGO maintains stable framing, consistent subject retention, and smooth motion while adhering to compositional intent.



Prompt: The camera, positioned at a low angle and steady, zooms out from a medium close-up while starting from the right back towards the middle center. This shot maintains a focused composition throughout the sequence.

Prompt: The camera pushes in from a medium shot, starting at eye level and positioned to the right front of the subject. Maintaining steady movement, it shifts focus up center while keeping the frame centered.

Prompt: The camera performs a dolly zoom in from an extreme long shot, starting at eye level and facing backward toward the middle center of the screen. Throughout the shot, the camera remains steady, maintaining a consistent focus on the subject.

Fig. 4: Additional qualitative comparison in Unity rendering and video generation. We compare CCD, DIRECTOR, GenDoP, and VERTIGO in both Unity-rendered previews and downstream video generation. VERTIGO produces more stable framing and stronger target retention in Unity, which transfers to more reliable composition and fewer visual failures in generated videos.

References

1. Bonatti, R., Wang, W., Ho, C., Ahuja, A., Gschwindt, M., Camci, E., Kayacan, E., Choudhury, S., Scherer, S.: Autonomous aerial cinematography in unstructured environments with learned artistic decision-making. Journal of Field Robotics **37**(4), 606–641 (2020)
2. Chen, Y., Rao, A., Jiang, X., Xiao, S., Ma, R., Wang, Z., Xiong, H., Dai, B.: Cinepregen: Camera controllable video previsualization via engine-powered diffusion. arXiv preprint arXiv:2408.17424 (2024)
3. Courant, R., Dufour, N., Wang, X., Christie, M., Kalogeiton, V.: E.t. the exceptional trajectories: Text-to-camera-trajectory generation with character awareness (2024), `https://arxiv.org/abs/2407.01516`
4. Courant, R., Wang, X., Loiseaux, D., Christie, M., Kalogeiton, V.: Pulp motion: Framing-aware multimodal camera and human motion generation. arXiv preprint arXiv:2510.05097 (2025)
5. Evin, I., Hämäläinen, P., Guckelsberger, C.: Cine-ai: Generating video game cutscenes in the style of human directors. Proceedings of the ACM on Human-Computer Interaction **6**(CHI PLAY), 1–23 (2022)
6. Galvane, Q.: Automatic cinematography and editing in virtual environments. Ph.D. thesis, Grenoble 1 UJF-Université Joseph Fourier (2015)
7. Gschwindt, M., Camci, E., Bonatti, R., Wang, W., Kayacan, E., Scherer, S.: Can a robot become a movie director? learning artistic principles for aerial cinematography. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1107–1114. IEEE (2019)
8. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
9. Hayashi, M., Inoue, S., Douke, M., Hamaguchi, N., Kaneko, H., Bachelder, S., Nakajima, M.: T2v: New technology of converting text to cg animation. ITE Transactions on Media Technology and Applications **2**(1), 74–81 (2014)
10. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
11. Hu, T., Zhang, J., Yi, R., Wang, Y., Huang, H., Weng, J., Wang, Y., Ma, L.: Motionmaster: Training-free camera motion transfer for video generation (2024)
12. Huang, C., Dang, Y., Chen, P., Yang, X., Cheng, K.T.: One-shot imitation drone filming of human motion videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5335–5348 (2021)
13. Huang, C., Lin, C.E., Yang, Z., Kong, Y., Chen, P., Yang, X., Cheng, K.T.: Learning to film from professional human motion videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4244–4253 (2019)
14. Jiang, H., Christie, M., Wang, X., Liu, L., Wang, B., Chen, B.: Camera keyframing with style and control. ACM Transactions on Graphics (TOG) **40**(6), 1–13 (2021)
15. Jiang, H., Wang, X., Christie, M., Liu, L., Chen, B.: Cinematographic camera diffusion model. In: Computer Graphics Forum. vol. 43, p. e15055. Wiley Online Library (2024)
16. Jiang, X., Rao, A., Wang, J., Lin, D., Dai, B.: Cinematic behavior transfer via nerf-based differentiable filming. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6723–6732 (2024)

17. Karakostas, I., Mademlis, I., Nikolaidis, N., Pitas, I.: Shot type constraints in uav cinematography for autonomous target tracking. Information Sciences **506**, 273–294 (2020)
18. Kuang, Z., Cai, S., He, H., Xu, Y., Li, H., Guibas, L.J., Wetzstein, G.: Collaborative video diffusion: Consistent multi-video generation with camera control. Advances in Neural Information Processing Systems **37**, 16240–16271 (2024)
19. Louarn, A., Christie, M., Lamarche, F.: Automated staging for virtual cinematography. In: Proceedings of the 11th ACM SIGGRAPH Conference on Motion, Interaction and Games. pp. 1–10 (2018)
20. Pueyo, P., Dendarieta, J., Montijano, E., Murillo, A.C., Schwager, M.: Cinempc: A fully autonomous drone cinematography system incorporating zoom, focus, pose, and scene composition. IEEE Transactions on Robotics **40**, 1740–1757 (2024)
21. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control (2025), `https://arxiv.org/abs/2503.03751`
22. Subramonyam, H., Li, W., Adar, E., Dontcheva, M.: Taketoons: Script-driven performance animation. In: Proceedings of the 31st annual ACM symposium on user interface software and technology. pp. 663–674 (2018)
23. Wang, B., Chen, X., Gadelha, M., Cheng, Z.: Frame in-n-out: Unbounded controllable image-to-video generation (2025), `https://arxiv.org/abs/2505.21491`
24. Wang, Q., Luo, Y., Shi, X., Jia, X., Lu, H., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation (2025), `https://arxiv.org/abs/2502.08639`
25. Wang, X., Courant, R., Christie, M., Kalogeiton, V.: Akira: Augmentation kit on rays for optical video generation. arXiv preprint arXiv:2412.14158 (2024)
26. Wang, X., Courant, R., Shi, J., Marchand, E., Christie, M.: Jaws: Just a wild shot for cinematic transfer in neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16933–16942 (2023)
27. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
28. Xie, C., Hemmi, I., Shishido, H., Kitahara, I.: Camera motion generation method based on performer's position for performance filming. In: 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE). pp. 957–960. IEEE (2023)
29. Xing, J., Mai, L., Ham, C., Huang, J., Mahapatra, A., Fu, C.W., Wong, T.T., Liu, F.: Motioncanvas: Cinematic shot design with controllable image-to-video generation (2025), `https://arxiv.org/abs/2502.04299`
30. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
31. Yu, Z., Wang, H., Katsaggelos, A.K., Ren, J.: A novel automatic content generation and optimization framework. IEEE Internet of Things Journal **10**(14), 12338–12351 (2023)
32. Yu, Z., Wu, X., Wang, H., Katsaggelos, A.K., Ren, J.: Automated adaptive cinematography for user interaction in open world. IEEE Transactions on Multimedia **26**, 6178–6190 (2023)
33. Yu, Z., Yu, C., Wang, H., Ren, J.: Enabling automatic cinematography with reinforcement learning. In: 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 103–108. IEEE (2022)

34. Yuan, Y., Wang, X., Wickremasinghe, T., Nadir, Z., Ma, B., Chan, S.H.: Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics (2025), https://arxiv.org/abs/2509.21309
35. Zhang, M., Wu, T., Tan, J., Liu, Z., Wetzstein, G., Lin, D.: Gendop: Auto-regressive camera trajectory generation as a director of photography (2025), https://arxiv.org/abs/2504.07083