VERTIGO: Visual Preference Optimization for Cinematic Camera Trajectory Generation

Mengtian Li[1,2], Yuwei Lu[1], Feifei Li[1], Chenqi Gan[1], Zhifeng Xie[1,2], and Xi Wang[3⋆]

[1] Shanghai Film Academy, Shanghai University
[2] Shanghai Engineering Research Center of Motion Picture Special Effects
[3] LIX, École Polytechnique, CNRS, IPP
http://vertigo.magic-lab.tech/

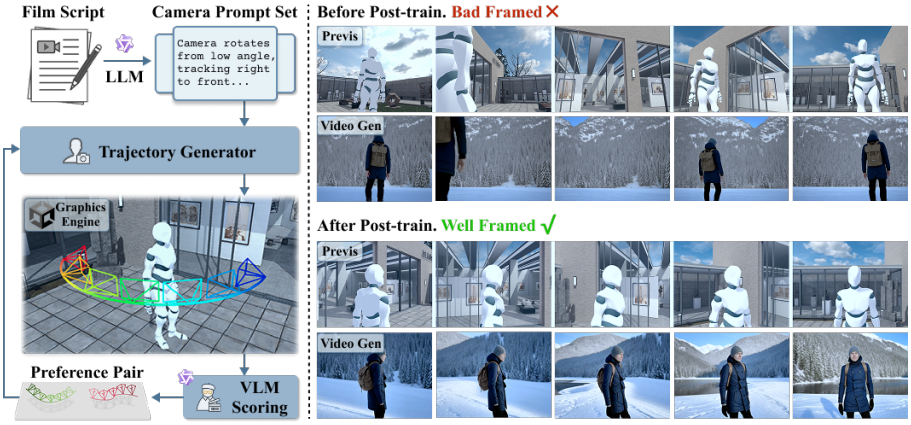

Fig. 1: Overview of VERTIGO. We present an integrated framework that converts film scripts into 3D camera trajectories refined via preference-based post-training. On the right, we show GenDoP results before and after post-training, demonstrating improved framing and quality in both graphics engine rendering and video generation.

Abstract. Cinematic camera control relies on a tight feedback loop between director and cinematographer, where camera motion and framing are continuously reviewed and refined. Recent generative camera systems can produce diverse, text-conditioned trajectories, but they lack this "director in the loop" and have no explicit supervision of whether a shot is visually desirable. This results in in-distribution camera motion but poor framing, off-screen characters, and undesirable visual aesthetics. In this paper, we introduce VERTIGO, the first framework for visual preference optimization of camera trajectory generators. Our framework leverages a real-time graphics engine (Unity) to render 2D visual previews from generated camera motion. A cinematically fine-tuned vision–language model then scores these previews using our proposed caption-based consistency similarity mechanism, which aligns renders with text prompts. This process provides the visual preference signals for Direct Preference Optimization (DPO) post-training. Both quantitative evaluations and user

⋆ Corresponding author.

studies on Unity renders and diffusion-based Camera-to-Video pipelines show consistent gains in condition adherence, framing quality, and perceptual realism. Notably, VERTIGO reduces the character off-screen rate from 38% to nearly 0% while preserving the geometric fidelity of camera motion. User study participants further prefer VERTIGO over baselines across composition, consistency, prompt adherence, and aesthetic quality, confirming the perceptual benefits of our visual preference post-training.

Keywords: Camera Trajectory Generation · Visual Preference Optimization · Computational Cinematography

1 Introduction

In the film industry, cinematography plays a pivotal role in transforming textual scripts into visual narratives through camera movement, framing, and composition, conveying emotional and aesthetic meaning to the audience on the silverscreen. Traditional filmmaking relies on close collaboration between the cinematographer, who interprets the director's instructions through camera movement, and the director, who supervises the on-screen composition from the monitor to decide whether a shot is kept or discarded.

With recent advances in computer graphics and computer vision, particularly in generative AI, many works in computational cinematography have begun exploring automated camera control systems that emulate the creativity of human cinematographers. Among recent developments, generative camera systems have shown great potential [2, 6, 21]. Most are capable of producing diverse and dynamic camera trajectories conditioned on textual prompts. However, unlike real-world collaboration between the cinematographer and the film director, these systems lack a "*director behind the monitor to validate or veto a shot*", *i.e.*, an **informative feedback mechanism** that guides the cinematographer in assessing whether the generated trajectories are aesthetically and narratively appropriate. In other words, simply generating in-distribution camera motions does not guarantee that the resulting shots exhibit good framing, spatial composition, or even the desired look-and-feel, a limitation also observed in recent work [8].

This collaborative generation paradigm naturally points to post-training techniques that incorporate reward or preference signals [18, 32, 33, 44, 65]. Such methods refine a pretrained generator by aligning it with downstream objectives through externally provided feedback, typically obtained from a reward model or a preference classifier that evaluates generated outputs. In our setting, this corresponds to equipping the camera generator with an additional evaluative signal capable of indicating whether a trajectory is desirable.

However, unlike most visual generation tasks, camera trajectories cannot be *directly* integrated into standard reward-based post-training frameworks. **Two main challenges arise**: (1) Unlike images or videos, a trajectory alone cannot be *directly* interpreted visually, whereas, in real filmmaking, directors make decisions by watching the monitor, not by inspecting raw camera paths; and (2) designing

a reliable evaluator for such a domain-specific task requires constructing robust metrics and supervision signals that can faithfully assess cinematic quality.

In this paper, we propose leveraging a real-time graphics engine (e.g., Unity) to generate *previews* (i.e., shot sequences rendered in the engine), which are then evaluated by a cinematically fine-tuned VLM using our caption-based consistency scoring mechanism that compares the semantic similarity between the original prompt and VLM-generated captions in embedding space. In this setup, the graphics engine functions as a 3D shooting stage, the camera generator plays the cinematographer, and the VLM-based evaluator serves as the *Director*. We then employ Direct Preference Optimization (DPO) [64] to post-train the camera trajectory generator on preference-ranked Unity previews, enabling it to better adhere to textual conditions, achieve more coherent framing, and produce trajectories with improved aesthetic look-and-feel.

Our method, VERTIGO, post-trains a GenDoP-style [6] camera generator on our *LenScript* dataset captioned with cinematic properties. Empirically, we observe consistent improvements in camera quality, particularly in condition adherence. To verify the visual improvement, we further evaluate VERTIGO using Unity rendering and diffusion-based Camera-to-Video generators, and report measurable gains in both rendered and generated video content. Most importantly, VERTIGO reduces the character off-screen rate from 38% to nearly 0%, while preserving, and in some cases even improving, geometric fidelity. User studies further confirm that VERTIGO achieves the highest preference rates in composition, consistency, prompt adherence, and aesthetic quality for both Unity-rendered previews and generated videos.

Our contributions are summarized as follows:

- We are the first to explore a complete **visual reward–based post-training framework** for camera trajectory generation. Our system leverages real-time rendered frames as visual rewards and employs an effective and robust VLM-based caption-based consistency scoring mechanism to produce preference signals for reinforcement-style post-training.
- We establish a reward system through our VLM-based caption-based consistency scoring mechanism, which computes latent semantic similarity between rendered shots and textual intentions. We proposed and compared multiple design choices and identified an effective strategy for using a VLM to evaluate trajectory quality from visual rewards, greatly improving framing quality.
- Through quantitative experiments and user evaluations, our DPO-trained camera generator produces improved trajectories in both Unity-rendered previews and camera-controlled video generation, demonstrating the effectiveness of the proposed reward and post-training framework.

2 Related Work

2.1 Automatic Virtual Cinematography

Recent advancements in virtual production, powered by real-time game engines [45, 46], have revolutionized filmmaking through dynamic previsualization.

Early computational approaches established foundational frameworks via geometric and rule-based camera control [6, 9, 13, 30, 43], while recent systems leverage predefined cinematic concepts [1, 41, 54] or LLM-driven multi-agent collaboration [57]. Optimization-based methods manage camera parameters and integrate aesthetic principles such as composition, viewpoint continuity, and action coherence [3, 10, 24, 35, 40, 59]. Learning-based approaches employ reinforcement learning [11, 61], deep networks [55, 60], and reference- or NeRF-based transfer of shooting patterns [16, 17, 20, 22, 52] to enhance trajectory quality, though these largely target aerial shots and lack fine-grained artistic control. Emerging diffusion-based methods [2, 21] generate plausible trajectories but struggle with multi-shot coherence and fine-grained control over shot scale, angle, and multi-segment planning; Pulp Motion [8] further incorporates framing-aware motion generation for improved immersion. Current methods fail to holistically integrate narrative context and cinematographic principles with visual aesthetic quality.

2.2 Controllable Video Generation

Controllable camera motion is essential for film-oriented video generation. While recent methods enable 2D-based motion control [12, 15, 58], they lack explicit 3D spatial modeling, causing perspective inconsistencies. 3D-aware approaches like MotionCtrl [53], CameraCtrl [3], and CVD [25] condition on simplified trajectories yet omit intrinsic parameters and cinematographic principles. AKiRa [5] models intrinsics but requires nontrivial trajectory acquisition. More recent systems further pursue 3D-informed or physically grounded synthesis [4, 47, 50, 56, 62], yet still treat camera control as a conditioning signal rather than an aesthetic objective. GenDoP [6] comes closest by autoregressively generating camera poses, yet existing frameworks universally neglect visual aesthetics and fail to unify text-driven narrative intent, cinematographic trajectory planning, and video generation under a cohesive aesthetic-aware system.

2.3 Reward-based Fine-tuning with VLMs

Reinforcement-learning-based post-training has become a prominent paradigm for aligning generative models with human preferences. In image and video generation, methods such as RLHF, RLAIF, and DPO have improved motion fidelity and text–visual consistency by distilling human-like preferences into diffusion models [1, 34, 64]. Works like Control-A-Video [4] further demonstrate reward feedback learning for controllable video diffusion, using visual reward signals derived from rendered frames. Self-rewarding and AI-generated preference signals further reduce dependence on costly labels and help mitigate reward hacking [28, 29, 39]. Recent Vision–Language Models (VLMs) also advance video understanding with finer temporal reasoning and narrative coherence [38, 49], and benchmark suites such as ShotBench and VEU-Bench [26, 31] probe cinematographic perception including shot scale, camera movement, and editing style. However, unlike images or videos where pixel-level visual feedback is readily available, camera trajectories are purely geometric sequences that carry no visual

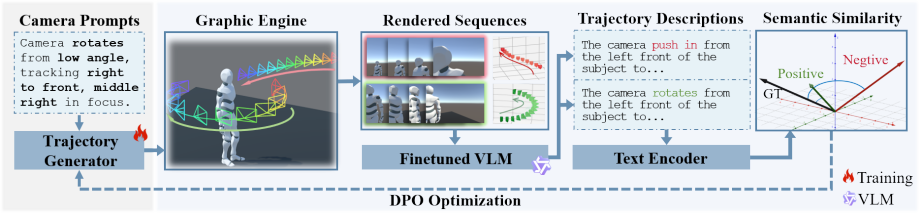

Fig. 2: Pipeline of VERTIGO. From a camera prompt, the generator produces 3D trajectories rendered into preview sequences by a graphics engine. A VLM performs inverse reasoning to caption the realized motion; the original prompt and generated caption are compared in latent space to derive preference scores for DPO post-training.

content on their own, making it impossible to directly derive visual reward signals. Applying VLM-derived rewards to camera-trajectory generation thus remains largely unexplored, requiring mechanisms that bridge trajectory space and visual space to enable preference-based optimization.

3 Method

Unlike prior approaches that focus primarily on geometric aspects of camera generation, our goal is to use fast, geometrically consistent rendering previews from a graphics engine (e.g., Unity) as *visual reward signals* to post-train camera generators in a reinforcement learning paradigm, so that the resulting camera motions lead to well-framed, aesthetically pleasing shots in both rendering-based and generation-based video production pipelines.

Central to our approach is a *visual feedback construction mechanism* that bridges the gap between trajectory-only camera generators and visual-quality assessment. Since raw trajectories carry no visual content, conventional post-training methods cannot directly evaluate or optimize their cinematic quality. Our mechanism leverages real-time graphics-engine rendering to materialize trajectories into preview sequences, and employs a VLM-based caption-based consistency scorer to derive preference signals, forming an automated, *render-in-the-loop* post-training pipeline. Concretely, the pipeline consists of three core technical components: (1) utilizing a pre-constructed Unity environment to render views generated by a pretrained camera generator; (2) employing caption-based consistency scoring within a VLM fine-tuned for camera motion classification to evaluate generated and augmented trajectories; (3) applying a DPO-style objective with positive and negative examples to post-train the camera generator. The complete pipeline is illustrated in Fig. 3, with detailed descriptions of each component provided in subsequent sections.

3.1 Trajectory Generator

We formalize the camera trajectory generation task as a conditional generation problem. Given a natural-language prompt p describing cinematographic intent

and an optional conditioning context c (e.g., scene layout, previous keyframe), the generator π_θ samples a camera trajectory τ from the learned distribution $\pi_\theta(\tau \mid p, c)$, parameterized by θ. During inference, the model produces temporally coherent camera sequences that respect both geometric constraints and narrative semantics encoded in p. This abstraction is compatible with various generation paradigms (e.g., diffusion or auto-regressive models) and serves as the basis for subsequent preference-based post-training.

Dataset: LenScript. Existing datasets [2, 6] reconstructed from videos lack fine-grained control over framing, shot composition, and focal properties, limiting their utility for training models that must reason about on-screen appearance. Since our trajectories are intended for previsualization and visual evaluation, we require training data that explicitly encodes cinematic intent at both geometric and visual levels. To this end, we procedurally construct **LenScript**, a trajectory dataset synthesized in Unity with precise annotations for both camera motion patterns and compositional attributes. Following the categorical structure of CCD [21], each trajectory is tagged with cinematic dimensions and paired with natural-language captions. Trajectories are synthesized in a local coordinate system and serialized in RealEstate10k-compatible format, supporting continuous focal length control. This dataset serves as the foundation for pretraining our trajectory generator, fine-tuning the VLM evaluator, and conducting ablation studies. Detailed dataset statistics are provided in the supplementary material.

Graphics Engine. Once the generator π_θ is trained, we deploy it in a Unity environment with multiple pre-constructed scenes for trajectory preview, editing, and render-sequence export. This real-time rendering enables preference-pair construction at negligible cost. Its trajectories are rendered into visual previews and passed to the VLM for preference scoring and post-training supervision. This engine-native pipeline ensures geometrically consistent, visually grounded evaluation of camera motion across diverse scene configurations.

3.2 Trajectory Preference in VLM

Motivation. To post-train the trajectory generator, we require clear and precise supervision signals. However, evaluating camera trajectories and their rendered previews is inherently challenging and multi-faceted, as factors such as framing, motion rhythm, and narrative intent rarely correlate with geometric or pixel-level errors. Unlike handcrafted heuristics that struggle with compositional nuances, vision-language models naturally ground visual cinematography in semantic space and enable holistic quality assessment through learned aesthetic representations.

More specifically, our preference scoring process proceeds as follows: given a textual prompt p, the trajectory generator samples a set of candidate camera motions $\{\tau_i\}_{i=1}^N$ using temperature and top-k sampling to promote exploration. Each trajectory τ_i is then rendered by the engine into a sequence of frames $\mathcal{I}_i = \{I_i^t\}_{t=1}^T$, along with a corresponding 3D path visualization. The goal is to evaluate how well each rendered preview $\mathcal{I}_i$ aligns with the textual intent expressed in p. We discuss three different approaches to achieve this:

(1) Tag-consistency scoring. We consider five canonical cinematography dimensions—the same categories as CCD's synthesis pipeline [21]: Camera Movement, Shot Scale, Shot Direction, Shot Angle, and Screen Property. For each dimension specified or implied by the textual prompt, the VLM judges whether the rendered sequence correctly satisfies it, yielding a count of matched dimensions in the range 0 to 5. We then discretize this count into a 10-level integer score $s_i^{\text{tag}} \in \{0, \ldots, 9\}$ (higher is better), directly reflecting how many of the five dimensions are correctly satisfied by the trajectory.
(2) Direct scalar regression. We prompt the VLM to reason about the five canonical dimensions and synthesize its judgments into a single aggregated match score $s_i^{\text{reg}} \in [0, 9]$ that reflects framing quality, motion rhythm, and compositional balance between $\mathcal{I}_i$ and p. This treats preference estimation as a unified scalar judgment rather than per-dimension verification. To obtain graded supervision and reduce labeling sparsity, we synthesize interpolated trajectories $\tilde{\tau} = \alpha\,\tau_i + (1-\alpha)\,\tau_j$ ($\alpha \in [0,1]$) between paths from different prompts, render the corresponding sequence $\tilde{\mathcal{I}}$, and assign $\hat{y} = \text{round}(9 \times \alpha)$ as the continuous supervision target. We then minimize an auxiliary RAFT-style token regression loss [36,37] that aligns the scalar digit with the auto-regressive token distribution:

$$\mathcal{L}_{\text{RAFT}} = \frac{1}{|\mathcal{D}|} \sum_{t \in \mathcal{T}_d} \left(\hat{y}_t - \sum_{d \in \mathcal{D}} p_t(d)\, d \right)^2, \tag{1}$$

where $p_t(d)$ is the predicted token distribution at decoding step t, $\mathcal{D}$ denotes the discretized score vocabulary, and $\hat{y}_t$ repeats the target score at each step to enforce consistent output across the generation sequence.
(3) Caption-based Consistency Scoring. Inspired by cycle-consistency principles in reward learning [2], we let the VLM perform inverse cinematic reasoning, producing a natural-language description $\hat{p}_i$ that summarizes the camera behavior in $\mathcal{I}_i$. We then encode both the original prompt p and the generated caption $\hat{p}_i$ using a dedicated semantic embedding model $\phi(\cdot)$ (a domain-adapted E5 model [48], fine-tuned on cinematographic captions from our LenScript dataset) and compute their cosine similarity in the latent space:

$$s_i^{\text{sem}} = \frac{\phi(p) \cdot \phi(\hat{p}_i)}{\|\phi(p)\|\,\|\phi(\hat{p}_i)\|}. \tag{2}$$

To enhance discriminative sensitivity on cinematographic captions, we apply lightweight LoRA contrastive fine-tuning on the embedding model (domain-adapted E5 [48]) using caption pairs from the LenScript dataset. This caption-based consistency scoring mechanism robustly measures semantic alignment between the textual intent and the realized camera motion, thereby providing grounded preference signals for DPO post-training.

Within each prompt-conditioned set, we apply one scoring method to rank trajectories and construct preference pairs (τ_i, τ_j) where $s_i > s_j$.
Choice of scoring strategy. Empirically, we find that both tag-consistency and direct regression suffer from inherent limitations when applied to cinematographic

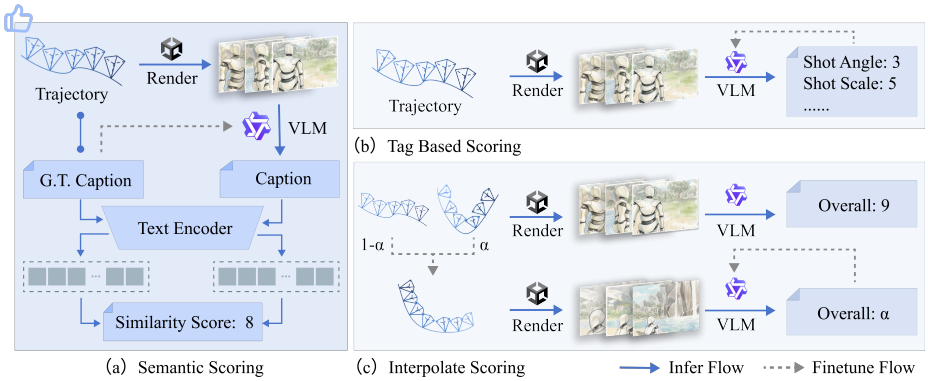

Fig. 3: Different VLM scoring and fine-tuning strategies of VERTIGO. We explore three preference scoring methods: (a) caption-based consistency scoring via latent-space similarity. (b) tag-consistency scoring; (c) direct scalar regression via RAFT-style fine-tuning on interpolated trajectories;

evaluation. VLMs struggle to provide fine-grained numerical judgments for camera trajectories, as discrete categorical scores (0–9) tend to cluster around a narrow range, losing discriminative power across visually similar but qualitatively different camera motions, which in turn hampers reliable construction of preference pairs. Moreover, these schemes exhibit weaker generalization across scenes and shot styles, and are brittle under prompt under specification. As a result, the regression variant tends to collapse to low-variance outputs (mode collapse), producing unreliable and poorly differentiated scores on Unity-rendered sequences with trajectory visualizations overlaid in 3D space. By contrast, caption-based consistency scoring provides richer signals through caption-space comparison, improving robustness to domain shift and enabling more diverse and stable preference ordering. Unless otherwise noted, we therefore adopt caption-based consistency scoring as the default scorer for all subsequent experiments.

VLM Fine-tuning for Cinematic Tasks. Off-the-shelf VLMs are trained on natural images and struggle with the domain gap introduced by our Unity-based renders, trajectory visualizations, and stylized lighting. This mismatch often leads the evaluator to misinterpret motion cues or framing, resulting in unstable preference judgments. To mitigate this, we start from **ShotBench** [31], a VLM already fine-tuned on large-scale cinematography data, and further adapt it to the Unity-rendered domain using lightweight LoRA fine-tuning. The model is trained on pairs of rendered sequences and natural-language camera captions, requiring no additional synthetic labels beyond the captions themselves.

3.3 Post-training for Trajectory Generation

Motivation. Supervised imitation alone does not align generation with nuanced cinematic preferences, as it treats all training samples equally and tends to overfit to seen prompts without distinguishing relative quality among plausible trajectories. Preference-based post-training addresses this by directly optimiz-

Tab. 1: Quantitative comparison of camera trajectory and video quality. We evaluate geometric fidelity using CLaTr-based metrics [2] and visual quality via VBench [27] metrics on both Unity-rendered sequences and controllable video generation. VERTIGO achieves competitive geometric performance while substantially improving visual metrics and target framing stability. Shaded indicates best performance.

Methods	Camera trajectory quality						Render Quality			Video Quality	
	FCD↓	CS↑	P↑	R↑	D↑	C↑	MisR.↓	Cons.↑	Aes.↑	Cons.↑	Aes.↑
CCD [21]	104.60	30.03	0.58	0.01	0.41	0.10	0.999	0.820	0.460	0.901	0.296
DIRECTOR [2]	26.28	32.20	0.87	0.29	0.79	0.44	0.803	0.892	0.464	0.908	0.293
GenDoP [6]	4.17	67.98	0.92	0.72	0.93	0.78	0.387	0.904	0.515	0.967	0.722
VERTIGO	4.22	68.40	0.92	0.73	0.93	0.78	0.008	0.908	0.518	0.969	0.735

ing pairwise quality comparisons, regularizing toward a reference policy, and improving cinematic realism without brittle reward engineering.

To this end, we employ Direct Preference Optimization (DPO) to post-train the trajectory generator with rendered-view preferences, using pairs automatically derived from the VLM-based scoring scheme. For each prompt p, candidate trajectories are ranked, and pairs (τ_i, τ_j) are formed where τ_i is preferred to τ_j. DPO optimizes a policy to increase the likelihood of preferred sequences relative to a frozen reference, without training an explicit reward model. Concretely, the relative reward is defined as $r_\theta(\tau \mid p) \coloneqq \log \frac{\pi_\theta(\tau|p)}{\pi_{\text{ref}}(\tau|p)}$, where π_θ is the learnable policy and π_{ref} is a pretrained reference. The corresponding loss:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(\tau_i,\tau_j)}[\log \sigma(\beta\,(r_\theta(\tau_i \mid p) - r_\theta(\tau_j \mid p)))]\,, \tag{3}$$

encourages correct pairwise ordering while implicitly regularizing toward π_{ref}, thereby preserving diversity and mitigating over-optimization.

Intuitively, this objective increases the log-odds of preferred trajectories relative to a fixed reference policy, serving as an implicit regularizer that limits distributional drift while improving alignment with the preference signal. The scalar β controls the pairwise margin, adjusting how strongly the policy diverges from its reference. In practice, optimizing this objective produces trajectories with higher preference scores: showing smoother motion, cleaner framing, and more stable spatial relations around targets.

4 Experiment

Setup. We evaluated the trajectory generation quality across geometric and visual dimensions. Geometric evaluation employed CLaTr-based [2] metrics to assess all methods on the full test set. For visual evaluation, we uniformly sample 1,000 trajectories from the test set under two conditions and measure their quality via VBench metrics [19]: (1) Direct rendering in four Unity scenes for visual content; (2) Video generation by transforming Unity-rendered videos using Wan 2.2 VACE [23], ensuring a comprehensive assessment of both geometric fidelity and downstream visual impact across rendering and generation paradigms.

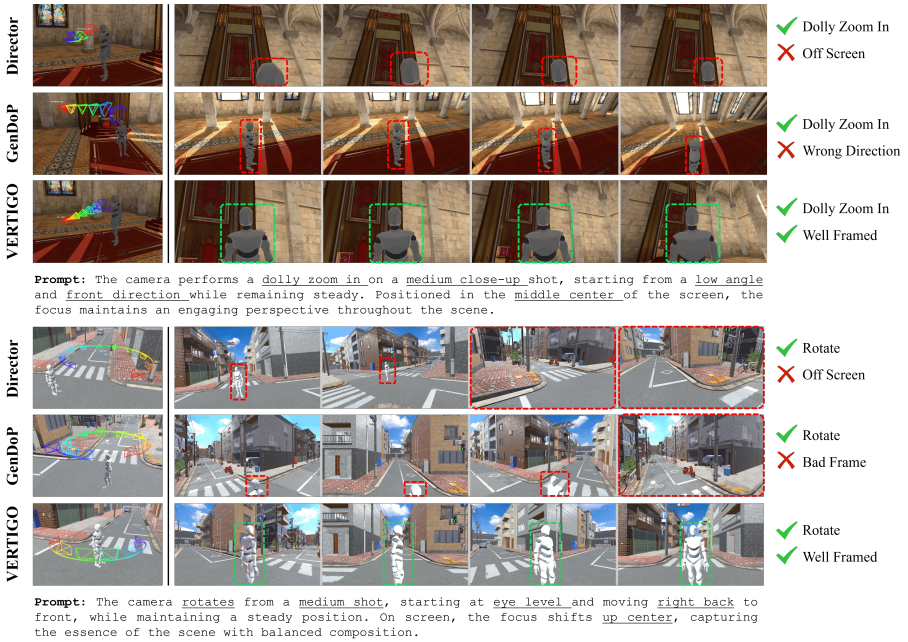

Fig. 4: Qualitative comparison of camera generators. VERTIGO accurately adheres to spatial composition instructions and maintains framing, whereas GenDoP and DIRECTOR misplace subjects and occasionally lose them.

Configs. We employ the Qwen 2.5-VL [49], fine-tuned on ShotBench [31], as the VLM for our preference scoring framework. For caption-based consistency scoring, we use a domain-adapted E5 [48] text embedding model, fine-tuned on cinematographic captions from our LenScript dataset. All experiments are conducted on two NVIDIA RTX 5880 Ada GPUs. Notably, our visual feedback construction pipeline—including real-time engine rendering, VLM-based preference scoring, and DPO optimization—is computationally lightweight. The graphics engine provides rendering that is orders of magnitude faster than diffusion-based video generators, and a compact VLM is sufficient for reliable cinematographic evaluation, enabling the entire post-training loop to converge efficiently without imposing significant computational overhead.

Metrics. We evaluate our method using two complementary groups of metrics measuring geometry-level trajectory quality and video-level perceptual quality.

– *Geometrical Metrics.* At the geometry level, we assess camera trajectory quality using a series of CLaTr-based metrics [2], including the Frechét CLaTr Distance (FCD), CLaTr-Score (CS), Precision (P), Recall (R), Density (D), and Coverage (C), which collectively quantify the fidelity and diversity of the generated camera trajectories with respect to the ground-truth distribution. To ensure a fair and consistent comparison across methods, we retrain the CLaTr baseline on both our LenScript dataset and the E.T. dataset.

Fig. 5: Qualitative comparison of video-to-video transfer results. We compare VACE-based video transfer on two trajectories. For the first trajectory, we show GenDoP vs. VERTIGO transfer results; for the second, we compare static and animated character renderings transferred under the same trajectory. VERTIGO produces more robust framing after transfer, while character animation does not affect transfer quality, as trajectories are defined in the subject's local coordinate space.

– *Visual Metrics*. At the video level, we further evaluate the perceptual impact of different trajectory generation models when applied to downstream video synthesis. Following the VBench [19], we measure consistency, and aesthetic, which respectively capture the temporal dynamic coherence, motion continuity, and perceptual appeal of the rendered video sequences. We also evaluate the Missing Rate (MisR) of the shooting target in Unity, which measures the fraction of frames where the shooting target is either off-screen or positioned within the outer 20% border region, so as to quantify the framing failure of the generated trajectory. Together, these metrics provide a comprehensive evaluation of both the structural realism of camera trajectories and their downstream effect on the final video generation quality.

4.1 Quantitative Results

VERTIGO demonstrates strong performance across both geometric and visual quality metrics. As shown in Tab. 1, we achieve competitive geometric metrics comparable to the current state-of-the-art trajectory generator in terms of competing FCD and enhanced CLaTr-Score, demonstrating that DPO post-training preserves distributional alignment with ground-truth camera motions. Notably, these geometric gains are obtained without additional geometric supervision.

Critically, while preserving geometric fidelity, our method substantially outperforms all baselines on visual quality metrics in both Unity-rendered and generated video settings. For example, VERTIGO reduces the Missing Rate

(MisR) to 0.008, compared to 0.387 for GenDoP, 0.803 for DIRECTOR, and 0.999 for CCD. In other words, our post-training nearly eliminates target framing failures. This directly reflects **the efficacy of visual-aware post-training** in preventing off-screen or edge-biased target placements that frequently occur in geometry-only approaches, and validates our core hypothesis that **rendered preview feedback is essential for stable target framing**. Beyond framing stability, our method also yields consistent gains on video-based metrics from VBench [27]. In the Unity rendering setting, VERTIGO improves both Consistency (0.908) and Aesthetic Quality (0.518), and in the video generation setting, it further achieves 0.969 Consistency and 0.735 Aesthetic Quality. These results show that our preference post-training with rendered previews successfully bridges geometric correctness and visual composition, producing trajectories that are structurally sound, visually compelling, and ready for downstream cinematic use in both CG and generative pipelines.

4.2 Qualitative Results

We present qualitative comparisons of trajectories generated by GenDoP, DIRECTOR and our method, rendered directly in Unity under identical prompts.
Prompt Adherence. Both GenDoP and DIRECTOR demonstrate reasonable compliance with basic motion instructions while exhibiting failures in handling compositional aspects that require spatial reasoning. For instance, when the user prompt specifies "positioned in the middle center of the screen", both baselines fail to correctly interpret this spatial constraint, misplacing the target. In contrast, VERTIGO accurately grounds the compositional instruction, precisely localizing the target within the specified screen region.
Visual Quality. Existing models often struggle to maintain focus on the target, sometimes even allowing it to disappear from the frame. This shortcoming stems from their geometric-only training paradigm, which neglects 2D visual feedback from the rendered image. VERTIGO overcomes this limitation through the proposed visual feedback construction mechanism, which enables render-in-the-loop post-training by materializing trajectories into preview sequences for VLM-based preference evaluation. As shown in Fig. 4, our method executes the correct motion pattern and faithfully preserves the "medium shot" composition, ensuring the subject remains properly framed as dictated by the prompt. We further evaluate video-to-video transfer quality by transforming Unity-rendered sequences using VACE under identical prompts. As shown in Fig. 5, we compare GenDoP and VERTIGO transfer results on one trajectory, and additionally compare static versus animated character transfers on a second trajectory. VERTIGO consistently maintains superior framing after transfer, while character animation has no effect on transfer quality, since trajectories are defined in the subject's local coordinate space. These results demonstrate that our visual preference signals transfer robustly to downstream video synthesis and generalize naturally across both static and dynamic subjects.
Animated Characters. We further evaluate our method on scenes with animated, non-static characters to verify compatibility with dynamic subjects. Since

Tab. 2: Ablation study. Starting from the GenDoP baseline, we compare three VLM-based preference scoring strategies for VERTIGO, evaluate generalization on animated characters, and ablate DPO β. All models are evaluated on Unity-rendered sequences using VBench metrics.

Metric	GenDoP	VERTIGO (Scoring Strategy)			VERTIGO (β ablation)			Anim. Characters	
		Tag	Interp.	Sem. (β=0.1)	β=0.01	β=0.5	β=0.9	GenDoP	Ours
MisR.↓	0.387	0.327	0.286	0.008	0.097	0.132	0.414	0.387	0.008
Cons.↑	0.904	0.902	0.906	0.908	0.897	0.905	0.901	0.876	0.902
Aes.↑	0.515	0.516	0.515	0.518	0.482	0.517	0.515	0.479	0.488

our trajectory is defined in a local coordinate system relative to the subject, camera poses are applied frame-by-frame to the subject's current position, naturally accommodating character animation. As illustrated in Fig. 5, VERTIGO maintains consistent framing and composition quality on animated characters, demonstrating that the visual feedback mechanism captures motion-agnostic framing principles rather than overfitting to static-scene configurations. Moreover, since our post-training operates solely on camera trajectories and is agnostic to rendering style, the learned preferences generalize naturally to diverse visual styles, from photorealistic rendering to stylized and artistic domains. Additional qualitative comparisons are provided in the supplementary materials.

Trajectory-Only Conditioning. We also examine conditioning a camera-based generator on the trajectory alone rather than the rendered preview. As shown in Suppl. Fig. 3, trajectory-only generation may fail to preserve the intended framing, whereas V2V (VACE) transfer retains the visual signal VERTIGO optimizes, justifying our V2V protocol.

4.3 Ablation Study

To validate the design choice and efficacy of our caption-based consistency scoring strategy in Sec. 3.2, we conduct an ablation study comparing all three candidate scoring methods. We post-train the camera generator with preference pairs constructed from different scoring methods, then evaluate the results on Unity-rendered sequences using VBench metrics.

Scoring Strategy. Our caption-based consistency scoring outperforms other strategies while avoiding mode collapse. As shown in Tab. 2, both tag-consistency and scalar regression exhibit significant limitations during DPO training. Discrete numerical scores lack sufficient granularity to distinguish subtle cinematographic variations, producing low-variance outputs that hamper reliable construction of preference pairs and fail to provide effective guidance for optimization. In contrast, caption-based consistency scoring leverages the rich semantic space of natural language captions, preserving fine-grained distinctions in camera behavior and framing intent. Our method achieves the best performance across all metrics, with dramatically reduced Missing Rate and improved consistency and aesthetic quality, validating that semantically grounded preference learning is essential for high-quality trajectory generation.

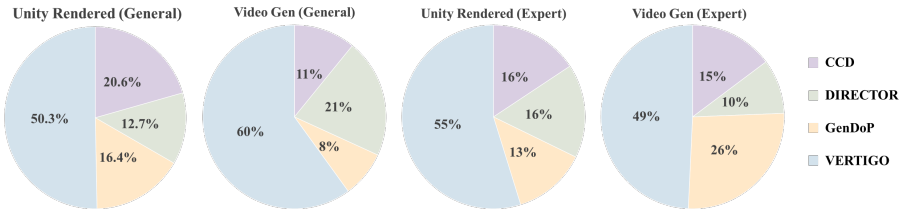

Fig. 6: User study best-of-4 results. Preference rates across evaluation dimensions for Unity rendering and video-to-video transfer (Wan 2.2 VACE). See supplementary for questionnaire details and additional results.

DPO Temperature β. We ablate the temperature parameter β, which controls how strongly the policy may diverge from the reference model. As shown in Tab. 2, β=0.1 yields the best overall trade-off, achieving the lowest Missing Rate while maintaining high consistency and aesthetic quality. Smaller values (e.g., β=0.01) under-regularize the policy, leading to degraded framing stability, while larger values ($\beta \geq 0.5$) over-constrain optimization and limit the model's capacity to learn effective visual preferences from the rendered feedback.

Animated Characters. To evaluate generalization beyond static scenes, we further test VERTIGO on sequences featuring animated characters with dynamic motion. As shown in Tab. 2, our method consistently improves framing quality over GenDoP on animated subjects, reducing the Missing Rate while improving both consistency and aesthetic scores. These results confirm that the visual preference signals acquired through our render-in-the-loop post-training transfer effectively to dynamic character scenarios, demonstrating that the learned framing behaviors are not tied to static-scene configurations.

Scorer Reliability. To confirm the VLM reward is not overfit to one rendering style, we evaluate the adapted scorer on Unity-default, Unity-prebuilt, and 3DGS scenes with PairAcc and Tag-F1 (Suppl. Tab. 2, Fig. 1). Our scorer stays consistently strong across all three domains, whereas the off-the-shelf VLM is unreliable, indicating cross-scene robustness.

4.4 User Study

To assess the perceived cinematic quality of generated trajectories, we conducted a human evaluation comparing VERTIGO against CCD, DIRECTOR, and GenDoP. We recruited 34 participants (11 experts with cinematography experience and 23 general participants) across two settings: (1) *Unity rendering*, directly reflecting trajectory quality, and (2) *video-to-video transfer* via Wan 2.2 VACE [23], assessing stylization fidelity and transfer resilience. Within each group of four videos (one per method, presented in randomized order), participants performed a *best-of-4 selection* task, choosing the best video along five perceptual dimensions per setting. Full questionnaire design and additional results (including full ranking and further comparisons) are provided in the supplementary materials.

As shown in Fig. 6, VERTIGO achieves the highest preference rates across all evaluation dimensions in both settings, with an overall win rate of 52.4%,

excelling in composition (51.8%) and instruction adherence (53.9%). These results confirm that our visual preference post-training effectively improves perceptual trajectory quality across both CG rendering and generative transfer modalities.

5 Conclusion

We present VERTIGO, a visual preference optimization framework that post-trains camera-trajectory generators using computer-graphics–rendered previews as visual reward signals. To tackle the challenging problem of evaluating trajectories through their resulting visual content, we introduce a VLM-based caption-based consistency scoring mechanism for assessing cinematographic quality, and systematically compare it against multiple alternative scoring designs through extensive ablations. Experiments show that our approach (1) dramatically reduces bad framings and improves prompt adherence while preserving high geometric fidelity, and (2) consistently enhances visual quality across both rendering-based and diffusion-based video generation pipelines. These improvements are corroborated by quantitative metrics and user studies. As the first framework to explicitly bridge geometric camera generation with shot visual outcomes, VERTIGO pushes computational cinematography toward practical, director-in-the-loop visual generation and broader real-world applications.

6 Limitations and Societal Impact

Limitations. Since the VLM serves as a preference annotator rather than the final generator, imperfect VLM perception constitutes an upper bound on preference quality: any annotation noise directly caps the quality of the DPO supervision. Moreover, our experiments only cover basic camera motion patterns, as these are amenable to precise annotation in synthetic environments; composite and long-take cinematographic motions remain unexplored. We also observe a failure mode of trajectory-only downstream generation: under a "rotate" trajectory, a camera-based video generator may distort the subject while attempting to keep it in-frame, indicating that trajectory conditioning alone does not always preserve the intended framing, motivating our V2V transfer protocol.
Societal Impact. Our synthetic data is procedurally generated and does not involve any copyrighted footage; nonetheless, downstream creative applications should still incorporate human-in-the-loop review to ensure the responsible and ethical use of automated cinematographic tools.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62402306), the NaturalScience Foundation of Shanghai (Grant No. 24ZR1422400, Grant No. 25ZR1401130), a Hi! Paris grant and fellowship. Computing resources by GENCI through access to the IDRIS High-Performance Computing facilities under allocations 2026-AD011014300R3.

References

1. Ahn, D., Choi, Y., Yu, Y., Kang, D., Choi, J.: Tuning large multimodal models for videos using reinforcement learning from ai feedback. In: ACL (2024)
2. Bahng, H., Chan, C., Durand, F., Isola, P.: Cycle consistency as reward: Learning image-text alignment without human preferences (2025), https://arxiv.org/abs/2506.02095
3. 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. J. Field Robot. **37**(4), 606–641 (2020)
4. Chen, W., Ji, Y., Wu, J., Wu, H., Xie, P., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning (2024), https://arxiv.org/abs/2305.13840
5. 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)
6. Christie, M., Olivier, P., Normand, J.M.: Camera control in computer graphics. In: Computer graphics forum. vol. 27, pp. 2197–2218. Wiley Online Library (2008)
7. Courant, R., Dufour, N., Wang, X., Christie, M., Kalogeiton, V.: E.t. the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In: ECCV (2024)
8. Courant, R., Wang, X., Loiseaux, D., Christie, M., Kalogeiton, V.: Pulp motion: Framing-aware multimodal camera and human motion generation. In: ICLR (2026)
9. Evin, I., Hämäläinen, P., Guckelsberger, C.: Cine-ai: Generating video game cutscenes in the style of human directors. ACM HCI **6**(CHI PLAY), 1–23 (2022)
10. Galvane, Q.: Automatic cinematography and editing in virtual environments. Ph.D. thesis, Grenoble 1 UJF-Université Joseph Fourier (2015)
11. 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: IROS. pp. 1107–1114. IEEE (2019)
12. 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. In: ICLR (2024)
13. Hayashi, M., Inoue, S., Douke, M., Hamaguchi, N., Kaneko, H., Bachelder, S., Nakajima, M.: T2v: New technology of converting text to cg animation. ITE MTAA **2**(1), 74–81 (2014)
14. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. In: ICLR (2025)
15. 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)
16. Huang, C., Dang, Y., Chen, P., Yang, X., Cheng, K.T.: One-shot imitation drone filming of human motion videos. IEEE TPAMI **44**(9), 5335–5348 (2021)
17. 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: CVPR. pp. 4244–4253 (2019)
18. Huang, Q., Chan, L., Liu, J., He, W., Jiang, H., Song, M., Song, J.: Patchdpo: Patch-level dpo for finetuning-free personalized image generation. In: CVPR. pp. 18369–18378 (2025)
19. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR (2024)

20. Jiang, H., Christie, M., Wang, X., Liu, L., Wang, B., Chen, B.: Camera keyframing with style and control. ACM TOG **40**(6), 1–13 (2021)
21. Jiang, H., Wang, X., Christie, M., Liu, L., Chen, B.: Cinematographic camera diffusion model. In: Comput. Graph. Forum. vol. 43, p. e15055. Wiley Online Library (2024)
22. Jiang, X., Rao, A., Wang, J., Lin, D., Dai, B.: Cinematic behavior transfer via nerf-based differentiable filming. In: CVPR. pp. 6723–6732 (2024)
23. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
24. Karakostas, I., Mademlis, I., Nikolaidis, N., Pitas, I.: Shot type constraints in uav cinematography for autonomous target tracking. Inf. Sci. **506**, 273–294 (2020)
25. 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. NeurIPS **37**, 16240–16271 (2024)
26. Li, B., Wu, Y., Lu, Y., Yu, J., Tang, L., Cao, J., Zhu, W., Sun, Y., Wu, J., Zhu, W.: Veu-bench: Towards comprehensive understanding of video editing. In: CVPR (2025)
27. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., Qiao, Y.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: CVPR (2024)
28. Li, Z., Yu, W., Huang, C., Liu, R., Liang, Z., Liu, F., Che, J., Yu, D., Boyd-Graber, J., Mi, H., Yu, D.: Self-rewarding vision-language model via reasoning decomposition (2025), https://arxiv.org/abs/2508.19652
29. Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., Ke, J., Dvijotham, K.D., Collins, K., Luo, Y., Li, Y., Kohlhoff, K.J., Ramachandran, D., Navalpakkam, V.: Rich human feedback for text-to-image generation. In: CVPR (2024)
30. Lino, C., Christie, M.: Intuitive and efficient camera control with the toric space. ACM TOG **34**(4), 1–12 (2015)
31. Liu, H., He, J., Jin, Y., Zheng, D., Dong, Y., Zhang, F., Huang, Z., He, Y., Li, Y., Chen, W., Qiao, Y., Ouyang, W., Zhao, S., Liu, Z.: Shotbench: Expert-level cinematic understanding in vision-language models. In: NeurIPS (2025)
32. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., et al.: Improving video generation with human feedback. In: NeurIPS (2025)
33. Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: CVPR. pp. 8009–8019 (2025)
34. Liu, R., Wu, H., Ziqiang, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: CVPR (2025)
35. Louarn, A., Christie, M., Lamarche, F.: Automated staging for virtual cinematography. In: ACM MIG. pp. 1–10 (2018)
36. Lukasik, M., Meng, Z., Narasimhan, H., Chang, Y.W., Menon, A.K., Yu, F., Kumar, S.: Better autoregressive regression with llms via regression-aware fine-tuning. In: ICLR (2025)
37. Lukasik, M., Narasimhan, H., Menon, A.K., Yu, F., Kumar, S.: Regression-aware inference with llms. In: Findings of EMNLP (2024)
38. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: EMNLP (2023)

39. Prabhudesai, M., Mendonca, R., Qin, Z., Fragkiadaki, K., Pathak, D.: Video diffusion alignment via reward gradients. In: ICLR (2025)
40. 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 TRO **40**, 1740–1757 (2024)
41. Rao, A., Jiang, X., Guo, Y., Xu, L., Yang, L., Jin, L., Lin, D., Dai, B.: Dynamic storyboard generation in an engine-based virtual environment for video production. In: SIGGRAPH Posters, pp. 1–2 (2023)
42. 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. In: CVPR (2025)
43. Subramonyam, H., Li, W., Adar, E., Dontcheva, M.: Taketoons: Script-driven performance animation. In: UIST. pp. 663–674 (2018)
44. Tong, C., Guo, Z., Zhang, R., Shan, W., Wei, X., Xing, Z., Li, H., Heng, P.A.: Delving into rl for image generation with cot: A study on dpo vs. grpo. arXiv preprint arXiv:2505.17017 (2025)
45. Unity: Real-time development platform (2025), `https://unity.com`
46. Unreal: We make the engine. you make it unreal. (2025), `https://www.unrealengine.com/`
47. 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`
48. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F.: Multilingual e5 text embeddings: A technical report. In: ACL (2024)
49. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution (2024), `https://arxiv.org/abs/2409.12191`
50. 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. In: ACM SIGGRAPH (2025)
51. Wang, X., Courant, R., Christie, M., Kalogeiton, V.: Akira: Augmentation kit on rays for optical video generation. In: CVPR (2025)
52. Wang, X., Courant, R., Shi, J., Marchand, E., Christie, M.: Jaws: Just a wild shot for cinematic transfer in neural radiance fields. In: CVPR. pp. 16933–16942 (2023)
53. 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. pp. 1–11 (2024)
54. Wei, Z., Wu, H., Zhang, L., Xu, X., Zheng, Y., Hui, P., Agrawala, M., Qu, H., Rao, A.: Cinevision: An interactive pre-visualization storyboard system for director-cinematographer collaboration (2025), `https://arxiv.org/abs/2507.20355`
55. Xie, C., Hemmi, I., Shishido, H., Kitahara, I.: Camera motion generation method based on performer's position for performance filming. In: GCCE. pp. 957–960. IEEE (2023)
56. 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. In: ACM SIGGRAPH (2025)
57. Xu, Z., Wang, J., Wang, L., Li, Z., Shi, S., Hu, B., Zhang, M.: Filmagent: Automating virtual film production through a multi-agent collaborative framework. In: SIGGRAPHASIA TC, pp. 1–4 (2024)

58. 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. pp. 1–12 (2024)
59. Yu, Z., Wang, H., Katsaggelos, A.K., Ren, J.: A novel automatic content generation and optimization framework. IEEE IoTJ **10**(14), 12338–12351 (2023)
60. Yu, Z., Wu, X., Wang, H., Katsaggelos, A.K., Ren, J.: Automated adaptive cinematography for user interaction in open world. IEEE TMM **26**, 6178–6190 (2023)
61. Yu, Z., Yu, C., Wang, H., Ren, J.: Enabling automatic cinematography with reinforcement learning. In: MIPR. pp. 103–108. IEEE (2022)
62. 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. In: ICLR (2026)
63. Zhang, M., Wu, T., Tan, J., Liu, Z., Wetzstein, G., Lin, D.: Gendop: Auto-regressive camera trajectory generation as a director of photography. In: ICCV (2025)
64. Zhang, R., Gui, L., Sun, Z., Feng, Y., Xu, K., Zhang, Y., Fu, D., Li, C., Hauptmann, A., Bisk, Y., Yang, Y.: Direct preference optimization of video large multimodal models from language model reward. In: NAACL (2025)
65. Zhu, Y., Wang, X., Lathuilière, S., Kalogeiton, V.: Soft-di [m] o: Improving one-step discrete image generation with soft embeddings. arXiv preprint arXiv:2509.22925 (2025)