NAVIGATIONAL COMPOSITIONAL GUIDANCE: LEARNING CONTENT-ADAPTIVE VIDEO TRAJECTORIES FOR PHOTOGRAPHY

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional image composition methods rely on basic geometric transformations and classical aesthetic rules for optimization. With the development of multi-modal models, numerous effective composition recommendation methods have been proposed recently. However, most existing methods merely provide static image-based composition references with simple text instructions, which cannot deliver refined and practical guidance for real shooting scenarios. Video-based composition guidance is more consistent with real-world dynamic rules due to video's inherent spatial-temporal continuity. Nevertheless, current video composition methods suffer from limited generalization for complex composition tasks, only supporting basic camera and perspective adjustments, and cannot meet fine-grained guidance demands. To tackle these drawbacks, this paper proposes a novel video model-based content-adaptive composition path recommendation algorithm, named Navigational Composition Guidance (NCG). It achieves complex-scene dynamic composition guidance and realizes autonomous navigation within the high-dimensional aesthetic manifold to search for optimal composition paths.

1 INTRODUCTION

The ubiquity of smartphones and digital cameras has transformed the capturing and documenting of life into an indispensable daily practice. Nonetheless, achieving masterful composition typically demands substantial expertise and professional experience. A professionally composed photograph requires a harmonious integration of various spatial elements, including the foreground, midground, and background. Therefore, a central objective of contemporary research is determining how to guide amateur users to achieve professional-grade results without the need for formal training.

Current traditional photographic guidance Deng et al. (2018); Guo et al. (2018); Hong et al. (2021b); Li et al. (2020b); Mai et al. (2016b); Wei et al. (2018) primarily relies on generating static image pairs to direct users, which often lacks explicit, actionable instructions. Video-based guidance makes intermediate compositional adjustments visible, helping users understand how to move the camera from an initial view toward a better composition. However, existing video-based guidance Yao et al. (2025) visualizes camera motion without explicitly modeling the aesthetic understanding needed for content-adaptive convergence toward aesthetic optima Esfandarani & Milanfar (2018). Evaluation also leaves a gap: frame-level aesthetics and overall video quality do not directly assess whether camera motion improves or degrades composition over time. These limitations motivate three challenges:

Scarcity of Expert Kinetic Data: The first challenge arises from the lack of specialized datasets. Current large-scale datasets predominantly consist of static paired images Zeng et al. (2022); Jia et al. (2022) or uncurated videos Murray et al. (2012), which lack the specific "before-and-after" mapping required for composition optimization. There is a profound deficit of high-quality video data that explicitly aligns amateur initial states with professional-grade camera trajectories, making it difficult for models to learn the priors of expert photographers.

Kinetic Realism and Photographic Prior Alignment: While infinite paths lead to aesthetic optima, most are photographically implausible. Naive optimizations often yield erratic, "unnatural" movements that violate professional conventions. The challenge lies in constraining navigational

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

exploration within a manifold of expert kinetic priors, ensuring the model favors structured maneuvers (e.g., trajectories A–D in Fig. 1) over arbitrary, mathematically-valid but visually-jarring "shortcuts".

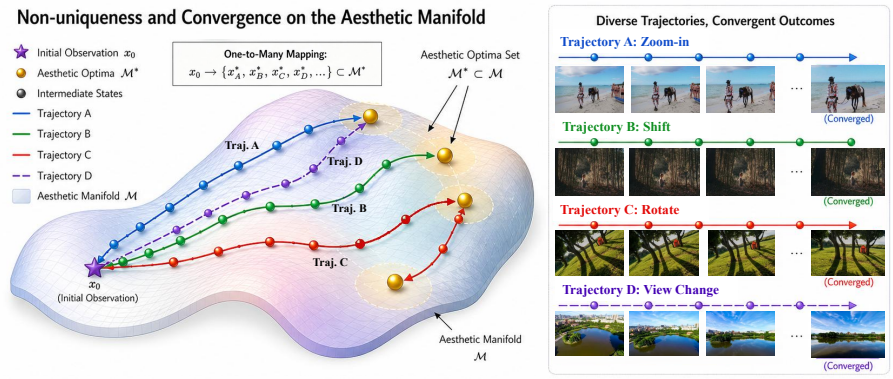

Figure 1: **Visualization of the non-convex aesthetic manifold and navigational convergence trajectories.** The surface illustrates the aesthetic scoring space where multiple valid camera trajectories (e.g., trajectories A-D) originate from a single observation x_0 and converge toward distinct aesthetic optima.

Solution Non-uniqueness and Content-Adaptive Convergence: As illustrated in Fig. 1, composition optimization is a one-to-many problem Huang et al. (2018); Zhu et al. (2017), with multiple valid optima for each initial state. The challenge is to adapt camera motion to visual content and settle into a quasi-static state near a favorable composition, avoiding premature stopping or continued motion after reaching it.

To address these challenges, we propose **Navigational Compositional Guidance (NCG)**, combining two-stage training for realistic, content-adaptive camera trajectories with a simulation-based data pipeline.

Our primary contributions are summarized as follows:

- We develop a streamlined NCG model that autonomously navigates the high-dimensional aesthetic manifold. Our approach eliminates explicit auxiliary commands and achieves implicit convergence via self-directed composition optimization, ultimately reaching state-of-the-art performance in video-based compositional guidance.
- We propose a two-stage training framework that facilitates the learning of expert kinetic priors, enabling the model to preserve realistic camera trajectories while ensuring stable, content-adaptive convergence on the aesthetic manifold.
- We introduce a camera-motion simulation pipeline and a high-quality dataset for learning photographic trajectories, together with Direction-Aware Photographic Evaluation (DAPE), a learned evaluator designed to assess whether camera motion improves or degrades composition over time.

2 Related Work

Photography Composition. Traditional photographic composition relies heavily on 2D image cropping, such as free-form Tu et al. (2020); Mai et al. (2016a); Li et al. (2018), subject-aware Hong et al. (2024b); Yang et al. (2023), or ratio-aware cropping Christensen & Vartakavi (2021). However, these methods are limited by the lack of 3D spatial manipulation, failing to provide effective guidance in complex cinematic scenarios. While PhotoFramer You et al. (2026) introduces composition optimization via image-text pairs, its reliance on semantic descriptions lacks intuitive visual cues, making it difficult for users to follow the intermediate compositional trajectory transitions. Similarly, although PPC Yao et al. (2025) achieves 3D composition optimization, it lacks compatibility with 2D operations and fails to adaptively adjust camera intensity based on content. Therefore, there is a critical need for a method that provides explicit, user-friendly composition intermediate guidance while seamlessly integrating both 2D and 3D camera maneuvers.

Aesthetic Evaluation. Aesthetic composition understanding is often an ill-posed problem without a single optimal solution, as aesthetic preferences vary significantly across individuals. Early research employed metrics such as FID Heusel et al. (2017), SWD Rabin et al. (2011), and CLIP scores Hessel et al. (2021) to simulate human visual perception; however, these remain implicit "black boxes" that yield only uninterpretable numerical values. With the advancement of Vision-Language Models (VLMs) Bai et al. (2025b;a), an increasing number of studies have focused on training models Liu et al. (2026) using curated human preference datasets. VLMs OpenAI (2024); Google (2025) can provide textual explanations alongside aesthetic scores, effectively simulating human preference evaluation with high interpretability. Furthermore, utilizing VLMs as reward models to optimize more powerful generative systems has emerged as a promising and encouraging direction.

3 Camera Movement Paradigm and Dataset Creation

3.1 Camera Movement Paradigm

Drawing inspiration from the kinetic maneuvers performed by professional photographers and formalize the photographic workflow as follows: First, the photographer selects a strategic physical location, which we define as the shooting point S_{point}. The handheld camera is then conceptualized as the camera point C_{point}. To optimize composition, the photographer operates through three fundamental degrees of freedom: translation, rotation, and focal length adjustment. We decompose these common camera movements into five primary categories He et al. (1996); Christie et al. (2008); Hou et al. (2024); You et al. (2026):

Zoom-in: Given an initial observation, this simulates an optical change in focal length. It corresponds to the focal scaling of the camera point C_{point}, conceptually analogous to a cropping operation in 2D space.

Shift: This simulates the lateral or vertical translation of the camera point C_{point} within the photographer's reach. This maneuver allows for the fine-tuning of compositional proportions and the re-balancing of visual weight within the frame.

Rotate: Simulates camera rotation to either rectify horizon misalignment or construct enhanced spatial geometric lines, corresponding to the rotational motion of the camera point C_{point}.

View Change: This simulates a strategic re-selection of the shooting point. Unlike localized adjustments of the shooting point S_{point}, a view change typically results in a significant transformation of the perspective, corresponding to the photographer physically relocating to a new vantage point.

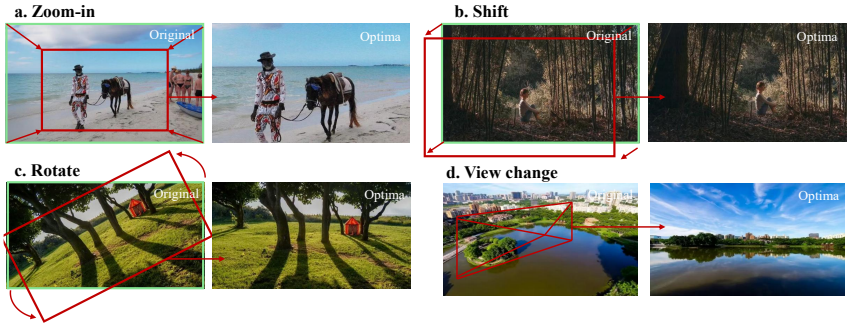

Figure 2: **Example of dataset construction.** Only the initial and final frames of the video sequences are displayed, with intermediate frames omitted. Red arrows indicate the direction of camera movement, while red boxes denote the target optimal frames.

Compound actions: Recognizing that compositional optimization often necessitates multi-axial adjustments, we define composite maneuvers aligned with professional photographic practices, such as *rotate+shift*, *view change+zoom-in*. These compound actions significantly extend the model's capability to navigate complex compositional scenarios and achieve sophisticated visual outcomes.

3.2 Dataset Creation

We have collected several widely-used public composition datasets, which feature crops and scores provided by professional photographers. These include GAIC Zeng et al. (2022), CPC Jia et al. (2022), FCDB Chen et al. (2017), SACD Hong et al. (2024a), FlickCrop Chen et al. (2016), CUHKCrop Yan et al. (2013), and Unsplash Unsplash. By integrating these datasets, we have established a foundation for the content-adaptive generation of composition-based videos.

3.2.1 Virtual Camera Model

For the three camera movements—zoom-in, shift, and rotate—the operations primarily take place within a 2D plane. This naturally allows for the simulation of camera motion directly on the image. We have designed a data production method based on virtual camera frame sampling. For a given image I, we define the virtual camera at time t as C_t, which is described by the following formula:

$$\mathbf{C}_t = [x_t, y_t, s_t, \theta_t]^T,$$

where (x_t, y_t) represent the coordinates of the virtual camera center on the original image (shift); s_t, the scaling factor (zoom) typically defined as the ratio between the original image and the camera frame; and θ_t, the rotation angle of the virtual camera (rotation).

We treat the image I as a function defined over a continuous domain $\Omega = [0, W] \times [0, H]$. The fusion process consists of establishing a mapping relationship from the output canvas coordinates $\mathbf{p}_{out} = (u, v)$ to the original image coordinates $\mathbf{p}_{in} = (x', y')$:

$$\mathbf{p}_{in} = \mathcal{T}_t(\mathbf{p}_{out}; \mathbf{C}_t)$$

where $\mathcal{T}_t$ is the affine transformation determined by the camera parameters $\mathbf{C}_t$ at time t. For a more detailed derivation of the mechanism and formulas, please refer to Appendix A.4.

To ensure high-quality composition in the final video frames, we rank images by professional photographer ratings within the dataset, selecting the top 20% as the target endpoints for the virtual camera recording. The parameters for $\mathbf{C}_t$ are then back-calculated using the corresponding cropping coordinates from the data annotations. Furthermore, to enhance the intensity of camera maneuvers, we employ Flux.1 to perform image outpainting on each image, with an expansion ratio k of 0.25. This expansion provides the virtual camera with a broader spatial range for motion simulation.

3.2.2 Data Construction

Dataset statistics. The overview of the dataset is presented in Table 1 and Fig. 3. Our dataset comprises a total of 23K raw video sequences, ensuring that the final frames all exhibit high-scoring compositions. Given the uneven distribution of certain data types, we performed random sampling on overrepresented categories to ensure a uniform data distribution. The training, testing and validation sets are partitioned with a ratio of $0.85 : 0.145 : 0.005$. Examples of the four primary camera movement are illustrated in Fig. 2.

Table 1: Dataset statistics.

	Zoom-in	Shift	Rotate	View change
Original	1116	4492	2780	446
sampling	1116	800	800	446
	Rotate+ Zoom-in	Rotate+ Shift	View change+ Zoom-in	Shift+ Zoom-in
Original	1050	5140	495	7864
sampling	1050	800	495	800

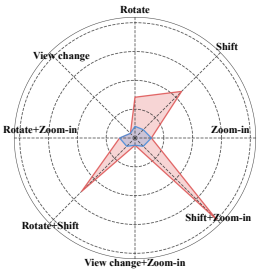

Figure 3: Visualization.

Zoom-in, shift, and rotate. For **zoom-in** data: First, we lock the virtual camera dimensions proportionally based on the aspect ratio of the final frame (high-scoring composition). Second, we filter out zoom-in videos with a scaling ratio lower than 1.2 to ensure that the camera movement characteristics are distinct. For **shift** data: We similarly lock the corresponding aspect ratio and

calculate the translation distance from the center coordinates of the final frame across eight directions, selecting only those videos with a displacement exceeding 100 pixels. For **rotate** data, starting from the final frame, we rotate the virtual camera both clockwise and counter-clockwise to the maximum angle α without exceeding the canvas boundary. Sequences with rotation angles α and $\alpha/2$ are selected and temporally reversed to generate the final video data.

View change. For **view change** data, as the perspective shift extends from 2D to 3D dimensions, simple virtual camera recording cannot yield plausible images. To address this, we developed a two-stage data production workflow: generating paired keyframes followed by temporal interpolation. First, we select professional-grade photographic images with the highest dataset scores to serve as the final frames. Subsequently, we employ Qwen-Image-Edit Bai et al. (2025b) to apply perspective rotations to these images, generating alternative viewpoints that function as the start frames. We logically assume these synthesized views to be compositionally sub-optimal, as the viewpoint shift inherently disrupts the precise geometric arrangements and aesthetic balance of the original expert-level compositions. In the next step, the Wan2.2 model Team (2025) is employed to perform interpolation between the start and end frames, obtaining the corresponding view change video data.

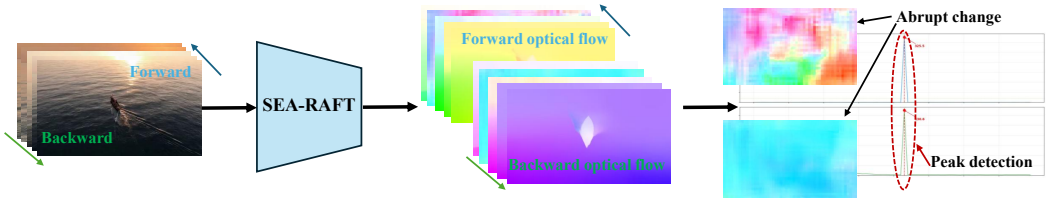

Figure 4: Bi-directional Flow Consistency Check.

Video Filtering. To mitigate artifacts like blur and frame-skipping common in generative data, we implement a dual-dimensional screening pipeline. First, we perform optical flow filtering using SEA-RAFT Wang et al. (2024b) to calculate bi-directional consistency errors. Sequences with abrupt error spikes, which usually indicate scene cuts, severe occlusions, or temporal inconsistencies, are excluded, as shown in Fig. 4. Second, we employ video quality filtering via PQA Yao et al. (2025) and Vbench 2.0 Huang et al. (2026), ranking videos across five dimensions: visual quality, motion smoothness, aesthetic appeal, temporal coherence, and background consistency. Our final dataset comprises the intersection of the top 50% from these rankings. We additionally conduct a human study to validate the aesthetic quality of the selected videos. See Appendix B for the detailed filtering workflow and human-study results.

Direction-aware evaluation. Our curated trajectories connect initial views to high-quality target compositions, providing a foundation for training DAPE to assess whether camera motion improves or degrades composition. DAPE and NCG are trained on disjoint sets of source images. Evaluator design, training, and validation are provided in Appendix D.

4 Methodology

4.1 Problem Formulation: Aesthetic Manifold Navigation

We define the image space as $\mathcal{X} \subset \mathbb{R}^{H\times W\times C}$. Within this high-dimensional space, there exists a low-dimensional aesthetic manifold $\mathcal{M}_A \subset \mathcal{X}$, representing the set of all visually "optimal" compositions.

Let $\mathcal{A} : \mathcal{X} \to \mathbb{R}$ be an aesthetic evaluation function that assigns a scalar score to any given frame x. The goal of Navigational Compositional Guidance (NCG) is to find a trajectory $\tau = \{x_0, x_1, \ldots, x_T\}$ such that the terminal state x_T resides in the set of local optima Ω on the manifold:

$$\Omega = \{x^* \in \mathcal{M}_A \mid \nabla\mathcal{A}(x^*) \to 0, \mathcal{A}(x^*) > \gamma\}$$

where γ is a high-aesthetic threshold. Given an initial amateur composition $x_0 \notin \mathcal{M}_A$, the optimization of x_0 is an ill-posed problem because Ω is non-convex and comprises multiple disjoint or degenerate local maxima. We model the trajectory generation as a conditional distribution $p(\tau \mid x_0)$. When solving the trajectory, the main challenge can be formulated as follows:

Existence of Multiple Attractors: For a single image x_0, there exist multiple distinct target states $\{x_T^{(1)}, x_T^{(2)}, \ldots, x_T^{(k)}\} \subset \Omega$. Each target corresponds to a different professional photographic intent.

Bifurcation of Trajectories: The mapping $\mathcal{F} : x_0 \to \tau$ is a one-to-many mapping. In the latent flow of the video model, this is expressed as:

$$p(\tau \mid x_0) = \sum_{i=1}^{k} \pi_i(x_0) p(\tau_i \mid x_0, c_i),$$

where c_i represents an implicit camera motion prior and π_i is the prior probability of choosing that specific maneuver. To solve the above issues, our model needs to learn a policy π_θ that maximizes the expected aesthetic return while maintaining smoothness of trajectory.

4.2 NCG Model

Our objective is to facilitate content-adaptive navigation within the aesthetic manifold, ensuring that trajectory endpoints converge toward the set of local optima Ω. To achieve this, we propose a two-stage training strategy. In the first stage, the model undergoes Supervised Fine-Tuning (SFT) using curated camera-motion videos to internalize professional photographic priors. This empowers the model to traverse diverse trajectories across the aesthetic manifold. In the second stage, we leverage Group Relative Policy Optimization (GRPO) to explore the multimodal aesthetic space. By comparing a group of sampled trajectories $\{\tau_1, \tau_2, \ldots, \tau_G\}$, the model reinforces those that successfully converge toward valid optima in Ω.

4.2.1 First Stage: Fine-Tuning

Multiple local attractors exist within the aesthetic manifold space. Our model is designed to explore this manifold by simulating the actual camera movements of a professional photographer. Due to the nature of one-to-many mapping, the model can theoretically traverse arbitrary paths to reach a destination. However, unconstrained trajectory generation often produces implausible motions, where the video model yields continuous transitions that are physically or artistically inconsistent with real-world cinematography. Consequently, the model should learn not only the target aesthetic state, but also realistic cinematic paths toward it.

To achieve this, we propose a progressive prompting strategy for video model fine-tuning. In the initial **strong supervision phase**, we train the model on explicit instruction-video pairs. During the **annealed substitution phase**, each specific prompt P_{spec} is stochastically replaced by a generic prompt P_{gen} with a time-varying probability θ_t that gradually increases from 0 to 1. Finally, in the **steady-state convergence phase**, the model is trained exclusively using the generic prompt P_{gen}.

Prompt design: We pre-defined templates for eight camera movements as strong constraints to guide initial trajectory convergence. To reduce the user's operational burden of specifying motion types, we gradually replace these specific templates with a unified prompt during later training stages. This forces the model to implicitly select the most appropriate trajectory based on visual context. See Appendix C for template details.

4.2.2 Second Stage: GRPO

After the first-stage fine-tuning, the model learns to generate realistic camera-motion trajectories within the aesthetic manifold. However, this supervised training mainly teaches the model how to follow plausible cinematic motions, rather than how to determine whether a trajectory actually approaches a local aesthetic optimum. As a result, the model may still continue moving, stop prematurely, or converge to suboptimal compositions when explicit trajectory supervision is unavailable.

To endow the model with adaptive convergence ability, we introduce a second-stage reinforcement learning phase based on GRPO, inspired by SAGE-GRPO Zheng et al. (2026). In this stage, the model is optimized to prefer trajectories that improve both visual quality and compositional aesthetics while preserving the cinematic motion priors learned in Stage 1. We use PQA Yao et al. (2025) as the reward oracle, which evaluates each generated compositional video along three dimensions: visual quality (VQ), motion quality (MQ), and compositional aesthetics (CA).

We synthesize the multi-dimensional rewards into a unified policy gradient objective. The optimization process couples the relative advantage of compositional trajectories with a dual-constraint regularization. For each generated trajectory $\tau^{(i)}$, the raw scores from our reward oracle are first Z-score normalized to ensure scale consistency across dimensions. The scalar reward $r^{(i)}$ is then computed as a weighted sum of VQ, MQ, and CA:

$$r^{(i)} = w_{\text{VQ}}\tilde{s}^{(i)}_{\text{VQ}} + w_{\text{MQ}}\tilde{s}^{(i)}_{\text{MQ}} + w_{\text{CA}}\tilde{s}^{(i)}_{\text{CA}},$$

where the $w_{\text{VQ}} : w_{\text{MQ}} : w_{\text{CA}} = 1 : 1 : 2$. Within the GRPO framework, the advantage $\hat{A}^{(i)}$ is estimated by normalizing the fused reward $r^{(i)}$ against the group mean. This advantage is then coupled with the clipped surrogate objective $\mathcal{L}_{\text{clip}}(\theta)$ to reinforce superior trajectories:

$$\mathcal{L}_{\text{clip}}(\theta) = \mathbb{E}_{i,t}\left[\min\left(\rho_t^{(i)}\hat{A}^{(i)}, \text{clip}(\rho_t^{(i)}, 1-\varepsilon, 1+\varepsilon)\hat{A}^{(i)}\right)\right],$$

where $\rho_t^{(i)}$ is the probability ratio between the current and reference policies. To maintain the kinetic priors acquired in Stage 1 and prevent mode collapse, we integrate a Dual-KL regularization term. The final loss $\mathcal{L}_{\text{loss}}(\theta)$ is defined as:

$$\mathcal{L}_{\text{loss}}(\theta) = -\mathcal{L}_{\text{clip}}(\theta) + \beta\left(\lambda_1\mathcal{L}_{\text{KL},\mu} + \lambda_2\mathcal{L}_{\text{KL,step}}\right)$$

where $\mathcal{L}_{\text{KL},\mu}$ penalizes velocity field deviations, and $\mathcal{L}_{\text{KL,step}}$ regularizes trajectory-wise distribution shifts. The coefficient β is adaptively updated via a PID-based controller to ensure the KL divergence remains within a stable trust region.

5 Experiments

Metrics. We quantitatively evaluate our work from three perspectives: viewpoint accuracy, video quality, and human preference scores. To assess viewpoint accuracy, we utilize CLIP-F Radford et al. (2021), PSNR Horé & Ziou (2010), SSIM Wang et al. (2004), and LPIPS Zhang et al. (2018) to measure image similarity between the initial and final frames. Additionally, we report image quality assessment via DeQA-Score You et al. (2025) and image aesthetic assessment via Q-Align Wu et al. (2024). For video quality assessment, we adopt VBench2.0 Huang et al. (2026) as our benchmarking metric. Finally, for human preference scoring, we conduct comparisons using the PQA model Yao et al. (2025). We further report DAPE scores to assess the direction of compositional change along camera trajectories, complementing frame-level aesthetics and overall video quality.

Implementation details. We follow the training configurations of HunyuanVideo Wu et al. (2025) and SAGE-GRPO Zheng et al. (2026) to develop our model. For both training stages, we employ LoRA to fine-tune the DiT (Diffusion Transformer) backbone. The model is trained using the AdamW optimizer on four NVIDIA H20 GPUs with a batch size of 1. For further hyperparameter details, please refer to Appendix E.

5.1 Results

Qualitative results. Visualizations of the four primary camera movement types are illustrated in Fig. 5, demonstrating the capability of NCG to autonomously select optimal motion paths and converge toward local optima. In the Zoom example, the model identifies a face in the lower-right corner and intentionally crops it to maintain focus on the core subject. For the Rotate case, the model detects a tilted frame and automatically applies a rightward rotation for alignment. In the Shift scenario, where the subject is partially cut off, the model pans the camera to rectify the positioning. Finally, in the View Change instance, the model adjusts the viewpoint to bypass foreground obstructions (e.g., trees), ensuring a clear line of sight to the subject.

Quantitative results. We compare our method with mainstream video models Seedance (2026); Alibaba Group (2026); Yao et al. (2025) and image-based models Su et al. (2024) for composition optimization, as shown in Tab. 2. The upper part reports video-level metrics, whereas the lower part reports metrics computed on initial–final frame pairs or image pairs. VBench2.0 Huang et al. (2026) evaluates six dimensions: subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. Meanwhile, PQA Yao et al. (2025) assesses



Figure 5: Qualitative results for the four camera movement. Green arrows and boxes illustrate the camera movement trajectories.

Visual Quality (VQ), Motion Quality (MQ), and Compositional Aesthetics (CA). Furthermore, our Direction-Aware Photographic Evaluation (DAPE) model assigns NCG the highest score of 0.9328, demonstrating its superior ability to generate directionally correct and aesthetically improving camera trajectories. We also conduct a human study to evaluate composition optimization; its design and additional quantitative comparisons are provided in Appendix F.

Table 2: Performance comparison of composition optimization.

	VBench2.0 Metrics						PQA Model Metrics			DAPE Metric
Model	Subj. Cons.	Back. Cons.	Mot. Smth.	Dyn. Deg.	Aes. Qual.	Img. Qual.	**VQ↑**	**MQ↑**	**CA↑**	**DAPE↑**
PPC	0.8870	0.9235	0.9887	0.2403	0.5381	63.069	0.3872	0.3677	0.3680	0.8631
Seedance	0.9292	0.9510	**0.9901**	0.8214	0.5507	**64.084**	0.3610	0.3472	0.3618	0.7816
HappyHorse	0.8668	0.9186	0.9799	**0.9697**	0.5406	63.495	0.3500	0.3353	0.3426	0.7380
NCG (Ours)	**0.9491**	**0.9530**	0.9892	0.8036	**0.5535**	63.388	**0.4947**	**0.4941**	**0.4828**	**0.9328**

Model	DeQA↑	Q-Align↑	CLIP-F↑	PSNR↑	SSIM↑	LPIPS↓	Human Study ↑
PPC	3.476	0.7226	0.8589	14.137	0.4378	0.5019	6.208
Seedance2.0	**3.861**	**0.8163**	0.8907	13.068	0.4269	0.5326	5.726
HappyHorse	3.381	0.7332	0.7861	10.720	0.4154	0.6896	4.556
S2CNet	3.352	0.7118	0.8868	**15.035**	**0.4379**	**0.4167**	6.316
NCG (Ours)	3.561	0.7374	**0.9020**	13.459	0.4270	0.5092	8.458

Camera movement types are also a primary focus of our study. To obtain precise camera trajectories from our videos, we employed DUSt3R Wang et al. (2024a) for pose estimation. Our dataset is categorized by motion type, with camera trajectories estimated separately for each category. Subsequently, we applied Out-of-Distribution detection Liu et al. (2020) to quantify the similarity between the inferred and reference trajectories; these quantitative results are summarized in Tab. 3. The characteristic distributions are visualized via energy mapping in Fig. 6. For a detailed explanation of the evaluation principles, please refer to Appendix G.

Table 3: Motion Trajectory Consistency Analysis.

Model	Metric	Value
NCG (Ours)	Legal Rate↑	**94.46%**
	OOD Rej.↓	**5.540%**
	Mot. Smo.↓	**1.6324**
	Avg. Cert.↑	**0.685**

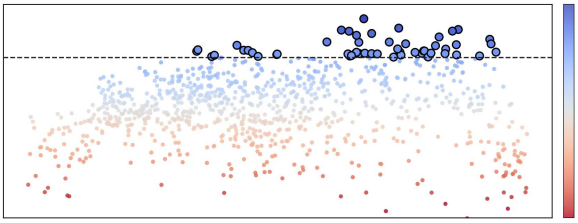

Figure 6: Visualization.

5.2 Ablation Studies and Discussions

Analysis of Gradient Interference. Due to the inherent one-to-many nature of the mapping within the compositional aesthetic manifold, a single initial state often corresponds to multiple valid endpoints. In the dataset, this manifests as identical images being paired with diverse camera trajectories. When a large batch size is employed, the simultaneous aggregation of gradients directing the model toward conflicting optima forces the optimization to gravitate toward a regressed mean state. This gradient interference results in significant visual artifacts, such as structural distortions and erratic camera jitter. Additional quantitative results, qualitative metrics and visual analyses are provided in Appendix H.

Analysis of training stage. We conduct an ablation study on our two-stage training strategy, with the evaluation results of the PQA model and OOD detection presented in Table 4. Our findings indicate that the first stage (SFT) effectively facilitates the convergence of trajectory plausibility. Subsequently, the second stage (GRPO) steers the model navigation toward the optimal compositional viewpoints. Detailed metrics for additional results are provided in Appendix H.

Table 4: Analysis of training stage.

	PQA Model Metrics			OOD Detection Metrics			
Model	VQ↑	MQ↑	CA↑	Legal Rate↑	OOD Rej.↓	Motion Smo.↓	Average Cer.↑
w/o SFT	0.4456	0.4515	0.4776	83.07%	16.93%	3.400	0.622
w/o GRPO	0.3946	0.3905	0.4349	92.03%	7.970%	3.589	0.644
NCG	**0.4947**	**0.4941**	**0.4828**	**94.46%**	**5.540%**	**1.632**	**0.685**

Aesthetic Convergence Analysis. We assess terminal settling and adaptation to initial composition quality. NCG reduces terminal motion by 34.78% and moves less on favorable inputs, preserving composition in 73% of cases while improving 82% of perturbed inputs. Together with the aesthetic evaluations, these results support content-adaptive adjustment and terminal settling (Appendix H).

6 Conclusion and Limitation

We propose the Navigational Compositional Guidance (NCG) model to enable autonomous convergence within complex aesthetic manifolds. NCG provides intuitive visual assistance through a two-stage training strategy that aligns model trajectories with professional photographic conventions. To support this, we develop a camera-motion simulation scheme and a high-quality dataset for generating expert-level kinetic video. This work offers a scalable solution for high-dimensional aesthetic optimization and opens new avenues for professional visual guidance. Current limitations include low data quality for large view change and convergence hurdles in composite maneuvers. Looking ahead, we plan to incorporate vision-language understanding to implicitly inject semantic knowledge, enhancing convergence for more sophisticated camera movements.

AI use statement

In this work, we used generative AI tools for polishing the English text and improving readability. We used generative models to construct synthetic training data, including image outpainting, viewpoint

editing, and video interpolation, as described in Section 3 and Appendices A–B. We have not used generative AI tools for generating core research ideas, writing the primary manuscript content, or generating scientific figures or plots. Additionally, we used generative AI tools for rephrasing certain sentences for better flow. We have reviewed all AI-assisted work. Specifically, all AI-polished text was carefully reviewed and heavily edited by the authors to ensure it accurately preserves our original scientific claims and technical meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies video-based guidance for photographic composition. Our data construction pipeline uses images from existing datasets and generates synthetic camera-motion sequences. The use and any redistribution of source images and derived data should respect the applicable licenses, usage restrictions, and privacy interests of depicted individuals. Generated views may contain hallucinated content or geometrically inaccurate details and should be treated as illustrative composition suggestions rather than faithful reconstructions of a scene or guarantees of physically feasible camera motion. Users should consider their surroundings and personal safety when following such guidance. Our evaluation includes human judgments of visual quality and composition. Aesthetic preferences are subjective and may vary across cultural backgrounds and photographic practices; the reported preferences and learned evaluation scores should therefore not be interpreted as universal standards of photographic quality.

REPRODUCIBILITY STATEMENT

We describe the camera-motion synthesis and data construction pipeline in Section 3 and Appendix A, with filtering procedures and dataset quality evaluations in Appendix B. Section 4 presents the two-stage training framework, while Appendices C and E provide the motion prompts and implementation details, including model configurations, optimization settings, and training infrastructure. Appendix D describes the construction, training, and independent direction-discrimination evaluation of DAPE. The evaluation metrics and experimental setup are introduced in Section 5, with additional comparisons and the human evaluation protocol in Appendix F, the trajectory evaluation procedure in Appendix G, and ablation and convergence analyses in Appendix H. Together, these descriptions document the procedures used to construct the data, train the models, and evaluate their outputs.

REFERENCES

Alibaba Group. Happyhorse ai. `https://happyhorse.app/`, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, and et al. Qwen3-vl technical report, 2025a. URL `https://arxiv.org/abs/2511.21631`.

Shuai Bai, Keqin Chen, Xuejing Liu, , and et al. Qwen2.5-vl technical report, 2025b. URL `https://arxiv.org/abs/2502.13923`.

Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. Automatic image cropping: A computational complexity study. In *CVPR*, pp. 507–515, Las Vegas, NV, USA, June 27-30 2016.

Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *Winter Conference on Applications of Computer Vision*, pp. 226–234, Santa Rosa, CA, USA, March 24-31 2017.

Casper L. Christensen and Aneesh Vartakavi. An experience-based direct generation approach to automatic image cropping. *IEEE Access*, 9:107600–107610, 2021.

Marc Christie, Patrick Olivier, and Jean-Marie Normand. Camera control in computer graphics. *Computer Graphics Forum*, 27(8):2197–2218, 2008.

Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *ACM Int. Conf. Multimedia*, pp. 870–878. ACM, 2018.

Tianxiang Du, Hulingxiao He, and Yuxin Peng. Venus: Benchmarking and empowering multimodal large language models for aesthetic guidance and cropping. In *CVPR*, pp. 37766–37776, Denver, CO, USA, June 3-7 2026.

Hossein Talebi Esfandarani and Peyman Milanfar. NIMA: neural image assessment. *IEEE Trans. Image Process.*, 27(8):3998–4011, 2018.

Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Guanjun Guo, Hanzi Wang, Chunhua Shen, Yan Yan, and Hong-Yuan Mark Liao. Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. *IEEE Trans. Multim.*, 20(8):2073–2085, 2018.

Li-wei He, Michael F. Cohen, and David Salesin. The virtual cinematographer: A paradigm for automatic real-time camera control and directing. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 217–224, New Orleans, LA, USA, August 4-9 1996.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pp. 6626–6637, Long Beach, CA, USA, December 2017.

Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *CVPR*, pp. 7057–7066, virtual, June 19-25 2021a.

Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *CVPR*, pp. 7057–7066, virtual, June 19-25 2021b.

James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning subject-aware cropping by outpainting professional photos. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *AAAI*, pp. 2175–2183, Vancouver, Canada, February 20-27 2024a.

James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning subject-aware cropping by outpainting professional photos. In *AAAI*, pp. 2175–2183, Vancouver, Canada, February 20-27 2024b.

Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, pp. 2366–2369. IEEE, 2010. doi: 10.1109/ICPR.2010.579.

Yunzhong Hou, Liang Zheng, and Philip Torr. Learning camera movement control from real-world drone videos, 2024. URL https://arxiv.org/abs/2412.09620.

Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pp. 179–196, Germany, September 8-14 2018.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(3):3268–3285, 2026.

Gengyun Jia, Huaibo Huang, Chaoyou Fu, and Ran He. Rethinking image cropping: Exploring diverse compositions from global views. In *CVPR*, pp. 2436–2445, New Orleans, LA, USA, June 18-24 2022.

Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-RL: aesthetics aware reinforcement learning for image cropping. In *CVPR*, Salt Lake City, UT, USA, June 18-22 2018.

Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *CVPR*, pp. 4212–4221, Seattle, WA, USA, June 13-19 2020a.

Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *CVPR*, pp. 4212–4221, Seattle, WA, USA, June 13-19 2020b.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di ZHANG, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *NeurIPS*, 2026.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, Red Hook, NY, USA, 2020. Curran Associates Inc.

Xiaoyu Liu, Ming Liu, Junyi Li, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Beyond image borders: Learning feature extrapolation for unbounded image composition. In *ICCV*, pp. 12977–12986, Paris, France, October 1-6 2023.

Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *CVPR*, pp. 497–506, Las Vegas, NV, USA, June 27-30 2016a.

Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *CVPR*, pp. 497–506, Las Vegas, NV, USA, June 27-30 2016b.

Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, pp. 2408–2415, Providence, RI, USA, June 16-21 2012.

OpenAI. Gpt-4o system card, 2024. URL `https://arxiv.org/abs/2410.21276`.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision*, pp. 435–446, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the ACM International Conference on Machine Learning (ICML)*, pp. 8748–8763, Virtual Event, July 18-24 2021.

Team Seedance. Seedance 2.0: Advancing video generation for world complexity, 2026. URL `https://arxiv.org/abs/2604.14148`.

Yukun Su, Yiwen Cao, Jingliang Deng, Fengyun Rao, and Qingyao Wu. Spatial-semantic collaborative cropping for user generated content. In *AAAI*, pp. 4988–4997, Vancouver, Canada, February 20-27 2024.

Wan Team. Wan: Open and advanced large video generation models. *arXiv preprint arXiv:2502.17347*, 2025.

Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *AAAI*, pp. 12104–12111, New York, NY, USA, February 7-12 2020.

Unsplash. The unsplash dataset. `https://unsplash.com/data`, 2020.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pp. 20697–20709, Seattle, WA, USA, June 16-22 2024a.

Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: simple, efficient, accurate RAFT for optical flow. In *ECCV*, pp. 36–54, Vienna, Austria, September 29-October 4 2024b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *CVPR*, pp. 5437–5446, Salt Lake City, UT, USA, June 18-22 2018.

Bing Wu, Chang Zou, Changlin Li, and et al. Hunyuanvideo 1.5 technical report, 2025. URL https://arxiv.org/abs/2511.18870.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the ACM International Conference on Machine Learning (ICML)*, pp. 54015–54029, Vienna, Austria, July 21-27 2024.

Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *CVPR*, pp. 971–978, Portland, OR, USA, June 23-28 2013.

Guo-Ye Yang, Wen-Yang Zhou, Yun Cai, Song-Hai Zhang, and Fang-Lue Zhang. Focusing on your subject: Deep subject-aware image composition recommendation networks. *Comput. Vis. Media*, 9(1):87–107, 2023.

Lujian Yao, Siming Zheng, Xinbin Yuan, Zhuoxuan Cai, Pu Wu, Jinwei Chen, Bo Li, and Peng-Tao Jiang. Photography perspective composition: Towards aesthetic perspective recommendation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *CVPR*, pp. 14483–14494, Nashville, TN, USA, June 11-15 2025.

Zhiyuan You, Ke Wang, He Zhang, Xin Cai, Jinjin Gu, Tianfan Xue, Chao Dong, and Zhoutong Zhang. Photoframer: Multi-modal image composition instruction. In *CVPR*, 2026.

Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3):1304–1319, 2022.

Ke Zhang, Tianyu Ding, Jiachen Jiang, Tianyi Chen, Ilya Zharkov, Vishal M. Patel, and Luming Liang. Procrop: Learning aesthetic image cropping from professional compositions. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *AAAI*, pp. 12600–12608, Singapore, January 20-27 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, Salt Lake City, UT, USA, June 18-22 2018. IEEE/CVF. doi: 10.1109/CVPR.2018.00068.

Mingzhe Zheng, Weijie Kong, Yue Wu, Dengyang Jiang, Yue Ma, Xuanhua He, Bin Lin, Kaixiong Gong, Zhao Zhong, Liefeng Bo, Qifeng Chen, and Harry Yang. Manifold-aware exploration for reinforcement learning in video generation, 2026. URL https://arxiv.org/abs/2603.21872.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, pp. 465–476, Long Beach, CA, USA, December 4-9 2017.

A Data Creation Overview

A.1 Analyzing Advantages of the Integrated Camera Motion Simulation Scheme

Our data synthesis workflow is characterized by its support for diverse motion types, streamlined configuration, and high degrees of freedom (DoF). It provides a comprehensive simulation framework across both 2D and 3D dimensions, facilitating rapid generation of 2D motion sequences. Compared to the current mainstream video-based camera motion construction methods, such as PPC, our approach offers superior usability and higher data fidelity.

Limitations of Current PPC Methods Yao et al. (2025) Existing PPC-based pipelines Yao et al. (2025) exhibit several significant drawbacks. First, they lack support for 2D data construction. Second, they rely heavily on complex 3D reconstruction to specify camera trajectories, which necessitates a sophisticated environment setup and incurs substantial computational overhead, as each frame requires individual 3D scene rendering. Furthermore, the visual quality is strictly bounded by the accuracy of the underlying 3D reconstruction, often leading to artifacts. Lastly, constrained by 3D scene geometry and hardware resources, these methods typically support only limited view angles and fail to simulate large-scale perspective transformations.

Advantages of Our Proposed Workflow In contrast, our method addresses these limitations through a dual-track strategy:

- **2D Motion Simulation:** By leveraging a virtual camera model, our scheme enables the rapid recording of diverse camera maneuvers. It allows for fine-grained control, where users can intuitively adjust the magnitude and intensity of movements via parametric tuning.
- **3D Motion Generation:** We have established a robust workflow based on two generative models integrated with a rigorous video filtering mechanism. This architecture supports high-throughput batch inference, ensuring both the diversity of viewpoints and the temporal consistency of the generated content.

A.2 Analysis of Camera Movement Design

To ensure our synthetic camera trajectories align with realistic professional photographic practices, we formalize our camera movement design into four atomic categories and four compound maneuvers. These designs are inspired by the fundamental principles of visual composition and the physical constraints of handheld photography.

Single Movements

- **Zoom-in:** This is the most prevalent technique for post-composition refinement. Mathematically, it is equivalent to a *cropping* operation during the post-processing stage. We prioritize this movement because it preserves the existing perspective relationships within the frame without introducing extraneous visual information from new viewpoints, making it the most reliable tool for emphasizing a subject.
- **Shift:** This maneuver simulates lateral or vertical camera translation within a localized workspace. It is strategically employed to rectify framing errors, such as when a primary subject is partially occluded by the frame boundary. By shifting the camera point C_{point}, photographers can either complete the visual representation of a subject or exclude distracting elements from the composition.
- **Rotate:** Rotational adjustments primarily address horizon misalignment caused by unstable handheld operation (i.e., camera tilt). Beyond error correction, intentional rotation is utilized to construct dynamic diagonal lines or "Dutch angles," which can enhance the geometric tension and aesthetic appeal of the scene compared to a standard horizontal alignment.
- **View Change:** This represents a high-level strategic relocation of the shooting point S_{point}. It is executed when the photographer's current coordinates cannot yield a viable composition due to physical obstacles or poor lighting angles. A view change implies a significant

transformation of the perspective, simulating the photographer physically moving to a superior vantage point.

Compound Actions In real-world scenarios, photographers often combine multiple degrees of freedom to achieve optimal framing. We design four representative compound actions:

- **Shift + Zoom-in:** This is the most common compound maneuver in professional workflows. Photographers typically perform a *Shift* first to center the subject or balance the visual weight, followed by a *Zoom-in* to eliminate peripheral distractions and finalize the composition scale.
- **Rotate + Zoom-in:** This serves as the primary compound correction for tilted frames. While the rotation rectifies the horizon, it often introduces empty or irregular triangular areas at the corners of the original rectangular sensor area; a synchronized *Zoom-in* (cropping) is thus essential to restore a clean, full-frame boundary.
- **Rotate + Shift:** This action is tailored for complex geometric realignment. It is employed when a photographer needs to both correct a tilt and simultaneously re-center a subject that has been displaced due to the rotation or the initial poor positioning. It reflects a high-precision adjustment of the camera's spatial orientation.
- **View Change + Zoom-in:** This simulates a "relocate-and-focus" workflow. After a photographer moves to a new strategic S_{point} to bypass obstacles or change the perspective, a *Zoom-in* is immediately applied to calibrate the specific field of view (FoV) for the target subject from the new vantage point.

A.3 Analysis of Unnatural Compound Actions

To ensure high-quality and realistic data, we explicitly exclude certain combinations of camera movements that deviate from professional photographic intuition. These "unnatural" actions often lead to visual instability or logical contradictions in the workflow. The following analysis justifies our selection:

- **Zoom-in + Shift:** This sequence implies finalizing the focal length before adjusting the camera's spatial position. In practice, a *Zoom-in* (especially at longer focal lengths) significantly amplifies the sensitivity of the frame to even minute camera displacements. Executing a *Shift* after zooming makes precise composition exceedingly difficult and often results in jittery or erratic visual trajectories that lack professional fluidness.
- **Zoom-in + Rotate:** This combination suggests that the photographer fixes the composition scale before correcting the horizon. This is counter-intuitive; establishing a stable horizontal reference is a fundamental prerequisite for effective framing. Reversing this order feels visually jarring and contradicts the standard "level-then-scale" operational hierarchy.
- **Zoom-in + View Change:** This maneuver represents a fundamental logical error in the photographic pipeline. It implies that a photographer would commit to a narrow field of view (FoV) before selecting a strategic vantage point. Professionally, one must first establish the optimal S_{point} to resolve perspective and occlusions before fine-tuning the internal focal parameters of the C_{point}.
- **View Change + {Shift, Rotate} or vice versa:** We observe a functional redundancy between *View Change* and localized adjustments (*Shift*/*Rotate*). While they differ in magnitude, they all serve to modify the camera's global or local spatial coordinates. In a realistic workflow, a decisive *View Change* is intended to reach a near-optimal position "in one go." Subsequent minor shifts or rotations are subsumed by the primary relocation, and treating them as distinct, simultaneous compound actions introduces unnecessary geometric noise.
- **Shift + Rotate:** Similar to the issues with *Rotate + Zoom-in* in reverse, performing a *Shift* before ensuring a level horizon is considered an inefficient workflow. Professional photographers prioritize "rectifying the canvas" (Rotate) as a baseline before executing lateral translations to balance subjects. Thus, we categorize this sequence as less representative of intentional artistic control.

A.4 Detailed Derivation of the Virtual Camera Transformation

To render a video frame at time t, we need to map each pixel coordinate $\mathbf{p}_{out} = [u, v, 1]^T$ in the target video canvas (output space) back to its corresponding coordinate $\mathbf{p}_{in} = [x', y', 1]^T$ in the original image I (input space). This inverse mapping ensures that we can sample the intensity values from the source image to fill the output buffer.

A.4.1 Coordinate System Alignment

We define the center of the output canvas as $(w/2, h/2)$, where w and h are the width and height of the video frame. The transformation $\mathcal{T}_t$ is composed of four sequential steps:

Normalization: Shift the output coordinate origin to the canvas center.

Zoom-in: Apply the scaling factor s_t

Rotation: Apply the rotation θ_t around the camera center.

Shift: Map the origin to the virtual camera's center (x_t, y_t) on the original image.

A.4.2 Matrix Decomposition

The composite transformation matrix $\mathbf{M}_t$ is uniquely determined by the virtual camera state $\mathbf{C}_t$ defined in the main text. It represents the mapping $\mathcal{T}_t$ and can be expressed as the product of elementary affine transformation matrices:

$$\mathbf{p}_{in} = \mathbf{M}_t \cdot \mathbf{p}_{out}$$

Substituting the components from $\mathbf{C}_t$, we have:

$$\mathbf{M}_t = \mathbf{T}(x_t, y_t) \cdot \mathbf{R}(\theta_t) \cdot \mathbf{S}(s_t) \cdot \mathbf{T}(-w/2, -h/2)$$

Where the elementary matrices are defined as:

Centering Translation: Shifts the output pixel to a local coordinate system centered at $(0, 0)$.

$$\mathbf{T}(-w/2, -h/2) = \begin{bmatrix} 1 & 0 & -w/2 \\ 0 & 1 & -h/2 \\ 0 & 0 & 1 \end{bmatrix}$$

Zoom Matrix: Adjusts the size of the sampled region. s_t represents the ratio of the source window to the target canvas.

$$\mathbf{S}(s_t) = \begin{bmatrix} s_t & 0 & 0 \\ 0 & s_t & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Source Translation: Positions the rotated and scaled frame at the virtual camera center (x_t, y_t) in the original image.

$$\mathbf{T}(x_t, y_t) = \begin{bmatrix} 1 & 0 & x_t \\ 0 & 1 & y_t \\ 0 & 0 & 1 \end{bmatrix}$$

Rotation Matrix: Rotates the sampling grid by θ_t.

A.4.3 Expanded Mapping Equation

By multiplying the matrices in order, we obtain the explicit coordinate mapping for $\mathbf{p}_{in} = [x', y', 1]^T$:

$$x' = s_t \cdot (u - w/2)\cos\theta_t - s_t \cdot (v - h/2)\sin\theta_t + x_t$$
$$y' = s_t \cdot (u - w/2)\sin\theta_t + s_t \cdot (v - h/2)\cos\theta_t + y_t$$

Image Resampling Since the calculated (x', y') are generally non-integer values, the final pixel value $V_t(u, v)$ is obtained via bilinear interpolation from the discrete pixel grid of image I:

$$V_t(u, v) = \sum_{i,j} I(i, j) \cdot \max(0, 1 - |x' - i|) \cdot \max(0, 1 - |y' - j|)$$

where (i, j) are the integer coordinates of the four pixels surrounding (x', y'). This ensures that the generated camera movements (zoom, shift, and rotation) remain spatially smooth and free of aliasing artifacts.

B Video Quality Filtering

To ensure training data quality, we apply a two-stage filtering pipeline that combines optical-flow-based motion filtering with multi-dimensional quality assessment via VBench.

B.1 Stage 1: Motion-based pre-filtering via Forward-Backward Optical Flow Consistency.

We first remove near-static videos using dense optical flow estimation (RAFT). Videos whose mean per-frame flow magnitude falls below a fixed threshold are discarded, producing a candidate set $\mathcal{V}_{\text{flow}}$.

To ensure the quality of training data collected from in-the-wild videos, we employ a forward-backward (FB) optical flow consistency check to automatically detect and remove video clips containing abrupt scene cuts, extreme motion blur, or other transient artifacts that would corrupt the learned representations.

Optical Flow Estimation. We use SEA-RAFT as the optical flow backbone. For a video with T frames $\{I_0, I_1, \ldots, I_{T-1}\}$, we process every consecutive frame pair (I_t, I_{t+1}). To avoid GPU memory overflow on high-resolution inputs, inference is optionally performed at a scaled resolution. Specifically, let $s \in \mathbb{Z}$ be a scale exponent (we use $s = -1$ for 1080p inputs). The model receives frames upsampled by 2^s and produces a flow field that is subsequently downsampled back to the original spatial resolution with a compensating scale correction:

$$\hat{\mathbf{F}} = \mathcal{M}(\uparrow_{2^s} I_t,\ \uparrow_{2^s} I_{t+1}), \qquad \mathbf{F}_{t\to t+1} = 2^{-s} \cdot \downarrow_{2^{-s}} \hat{\mathbf{F}}, \tag{1}$$

where $\uparrow_r$ and $\downarrow_r$ denote bilinear interpolation at scale factor r, and $\mathcal{M}$ is the flow network. The scalar 2^{-s} rescales the flow *values* to remain in the original pixel coordinate space after the spatial downsampling. The backward flow $\mathbf{F}_{t+1\to t}$ is computed analogously by swapping the input frame order.

Backward-Flow Warping. Given the forward flow $\mathbf{F}_{t\to t+1} \in \mathbb{R}^{H\times W\times 2}$ and the backward flow $\mathbf{F}_{t+1\to t} \in \mathbb{R}^{H\times W\times 2}$, let $\mathbf{p} = (x, y)^\top$ denote an integer pixel coordinate in frame I_t. The forward flow maps $\mathbf{p}$ to its estimated correspondence in I_{t+1}:

$$\mathbf{p}' = \mathbf{p} + \mathbf{F}_{t\to t+1}(\mathbf{p}). \tag{2}$$

Because $\mathbf{p}'$ is generally a non-integer location, we obtain the backward flow at $\mathbf{p}'$ via bilinear interpolation:

$$\widetilde{\mathbf{F}}_{t+1\to t}(\mathbf{p}) \ = \ \mathbf{F}_{t+1\to t}(\mathbf{p}') \ = \ \mathcal{B}(\mathbf{F}_{t+1\to t},\ \mathbf{p} + \mathbf{F}_{t\to t+1}(\mathbf{p})), \tag{3}$$

where $\mathcal{B}(\mathbf{F}, \mathbf{q})$ denotes bilinear sampling of field $\mathbf{F}$ at continuous position $\mathbf{q}$, with boundary pixels replicated for out-of-frame coordinates. In practice, Eq. equation 3 is implemented as a single `F.grid_sample` call on the GPU.

Per-Frame FB Consistency Error. Under ideal conditions (dense, non-occluded scene with perfectly estimated flow), the round-trip displacement from $\mathbf{p}$ back to itself should be zero:

$$\mathbf{F}_{t\to t+1}(\mathbf{p}) \;+\; \widetilde{\mathbf{F}}_{t+1\to t}(\mathbf{p}) \;=\; \mathbf{0}. \tag{4}$$

We define the *pixel-wise FB consistency error* as the ℓ_2 magnitude of the residual displacement:

$$\varepsilon_t(\mathbf{p}) \;=\; \left\| \mathbf{F}_{t\to t+1}(\mathbf{p}) \;+\; \widetilde{\mathbf{F}}_{t+1\to t}(\mathbf{p}) \right\|_2, \tag{5}$$

measured in pixels. Large $\varepsilon_t(\mathbf{p})$ indicates that the optical flow is mutually inconsistent at pixel $\mathbf{p}$, which typically arises from motion boundaries, occlusions, or — importantly for our filtering goal — scene discontinuities such as hard cuts or severe motion blur. From the pixel-wise map we derive two scalar statistics for the frame pair (I_t, I_{t+1}):

$$e_t^{\text{mean}} = \frac{1}{HW} \sum_{\mathbf{p}} \varepsilon_t(\mathbf{p}), \tag{6}$$

$$e_t^{\max} = \max_{\mathbf{p}} \; \varepsilon_t(\mathbf{p}). \tag{7}$$

Video-Level Abrupt Change Detection. For a video clip of T frames, the per-pair mean errors $\{e_0^{\text{mean}}, e_1^{\text{mean}}, \ldots, e_{T-2}^{\text{mean}}\}$ constitute a temporal profile of scene consistency. We define two complementary criteria for flagging a frame pair as an *abrupt transition*:

Absolute threshold: A frame pair is marked abrupt if its mean FB error exceeds an absolute bound τ_{abs} (in pixels):

$$\mathcal{A}_t^{\text{abs}} \;=\; \mathbf{1}[e_t^{\text{mean}} > \tau_{\text{abs}}]\,. \tag{8}$$

Relative threshold: To handle videos with globally high motion, we additionally compare each frame pair against the temporal median of the clip:

$$\mathcal{A}_t^{\text{rel}} \;=\; \mathbf{1}[e_t^{\text{mean}} > \tau_{\text{ratio}} \cdot \text{median}_{t'}(e_{t'}^{\text{mean}})]\,. \tag{9}$$

The denominator is stabilized with a small constant $\epsilon = 10^{-6}$ to prevent division by zero in static scenes. A frame pair (I_t, I_{t+1}) is deemed abrupt if either criterion is satisfied:

$$\mathcal{A}_t \;=\; \mathcal{A}_t^{\text{abs}} \;\vee\; \mathcal{A}_t^{\text{rel}}. \tag{10}$$

A video clip is then classified as containing an abrupt change if any frame pair is flagged:

$$\text{HasAbrupt} \;=\; \bigvee_{t=0}^{T-2} \mathcal{A}_t. \tag{11}$$

We use $\tau_{\text{abs}} = 20.0$ pixels and $\tau_{\text{ratio}} = 10.0$ throughout all experiments. The dual-threshold design is intentional: τ_{abs} catches gross failures regardless of video content, while τ_{ratio} catches within-video anomalies (e.g., a single hard cut embedded in an otherwise smooth clip).

Summary Statistics Recorded per Video. For each video, the following statistics are written to a `fb_stats.json` file for downstream analysis:

- Per-frame: e_t^{mean}, $e_t^{\max}$, and the mean and maximum magnitude of the forward flow $\|\mathbf{F}_{t\to t+1}(\mathbf{p})\|_2$.
- Video-level: $\max_t e_t^{\text{mean}}$, $\text{median}_t\, e_t^{\text{mean}}$, $\bar{e}^{\text{mean}}$ (temporal mean), the max-to-median ratio $\rho = \max_t e_t^{\text{mean}} / (\text{median}_t\, e_t^{\text{mean}} + \epsilon)$, the boolean flag HasAbrupt, and the list of flagged frame-pair identifiers.

Parallel Processing. Given a corpus of N videos, the pipeline is parallelised across K GPUs by uniformly partitioning the video list into K non-overlapping contiguous shards of size $\lceil N/K \rceil$. Each GPU runs an independent Python process that loads its own copy of the model weights, iterates over its shard, and appends completed video paths to a shared `done_videos.txt` log. Crash recovery is handled by checking this log at the start of each process, so partially completed runs can be resumed without reprocessing already-finished videos.

B.2 Stage 2: Multi-dimensional quality filtering.

We evaluate every candidate video with two complementary quality assessors: the VBench benchmark and a Perceptual Quality Assessment (PQA) model. VBench covers six dimensions of video quality:

- **Subject Consistency**: measures the temporal coherence of the primary subject's appearance across frames, penalizing identity drift or shape inconsistency caused by generation artifacts.
- **Background Consistency**: measures the stability of background regions over time, filtering videos with flickering or incoherent scene elements.
- **Motion Smoothness**: evaluates the naturalness of motion trajectories, identifying videos with abrupt or jittery frame transitions.
- **Dynamic Degree**: quantifies the overall magnitude of motion in the video, ensuring retained videos contain meaningful temporal dynamics rather than trivially static content.
- **Aesthetic Quality**: assesses perceptual visual quality using learned aesthetic features, capturing composition, color harmony, and overall visual appeal.
- **Imaging Quality**: measures low-level technical quality such as sharpness, noise level, and the absence of compression artifacts.

The PQA model assesses three distinct dimensions:

- **Visual Quality (VQ)**: evaluates frame-level perceptual fidelity, including clarity, color accuracy, and the absence of visual distortions.
- **Motion Quality (MQ)**: assesses the temporal coherence and plausibility of motion, penalizing unnatural acceleration, ghosting, or motion blur artifacts.
- **Compositional Aesthetics (CA)**: captures higher-level aesthetic properties of the visual layout, including spatial balance, subject framing, and scene composition.

Let $\mathcal{V}$ denote the candidate set after Stage 1. Define the joint dimension set $\mathcal{D} = \mathcal{D}_{\text{VBench}} \cup \mathcal{D}_{\text{PQA}}$, where $\mathcal{D}_{\text{VBench}}$ comprises the six VBench dimensions and $\mathcal{D}_{\text{PQA}} = \{\text{VQ}, \text{MQ}, \text{CA}\}$. For each dimension $d \in \mathcal{D}$, we retain the top-τ fraction of videos by their respective scores:

$$\mathcal{V}_d = \left\{ v \in \mathcal{V} \;\middle|\; s_d(v) \geq \text{Quantile}_{1-\tau}\left(\{s_d(u)\}_{u \in \mathcal{V}}\right) \right\}, \quad d \in \mathcal{D}. \tag{12}$$

The final training set is the intersection across all nine per-dimension sets:

$$\mathcal{V}_{\text{final}} = \bigcap_{d \in \mathcal{D}} \mathcal{V}_d. \tag{13}$$

This conjunction criterion ensures that retained videos simultaneously satisfy high-quality thresholds across *every* dimension from both assessors, rather than merely on average, providing complementary coverage of low-level technical quality, motion fidelity, and compositional aesthetics.

B.3 Human Study of Synthetic Dataset Quality

To assess the perceptual quality of our synthetic camera-motion dataset, we conduct a human study with 6 participants from three groups: two computer vision researchers, two photography experts, and two general users. We randomly sample 30 videos from each of five camera-motion categories: zoom-in, shift, rotation, view change, and compound actions, resulting in 150 videos in total. The compound-action category comprises trajectories involving multiple sequential camera movements. Each participant independently evaluates every video along four dimensions: visual quality, motion quality, compositional improvement, and final-frame composition quality. Visual quality measures the overall fidelity of the generated content, while motion quality evaluates the smoothness and physical plausibility of the camera movement. Compositional improvement measures whether the trajectory progressively improves the composition, and final-frame composition quality assesses the aesthetic quality of the resulting composition. Each dimension is rated on a five-point Likert scale, ranging from 1 (*poor*) to 5 (*excellent*). As summarized in Table 5, our synthetic videos receive consistently favorable ratings across all four dimensions, demonstrating that the constructed dataset exhibits high perceptual quality, plausible camera motion, and meaningful compositional improvement.

Table 5: Human evaluation of our synthetic camera-motion dataset. Scores are reported as mean on a five-point Likert scale, where higher values indicate better performance.

Motion Type	Visual Quality	Motion Quality	Compositional Improvement	Final-frame Composition
Zoom-in	4.8	4.9	4.6	4.5
Shift	4.8	4.9	4.8	4.6
Rotation	4.7	4.9	4.9	4.6
View change	4.8	4.6	4.6	4.5
Compound actions	4.7	4.7	4.8	4.6
Overall	4.76	4.80	4.74	4.56

B.4 Quantitative Evaluation of Synthetic Dataset Quality

In addition to the human study, we quantitatively evaluate our synthetic dataset using the same metrics and evaluation protocol as in Table 2. We report results separately for five camera-motion categories: zoom-in, shift, rotation, view change, and compound actions.

Video-level evaluation. We use VBench 2.0 to assess subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. PQA evaluates visual quality (VQ), motion quality (MQ), and compositional aesthetics (CA), while DAPE assesses whether camera trajectories produce aesthetically improving compositional transitions. The results are summarized in Table 6.

Image-level evaluation. We additionally report DeQA, Q-Align, CLIP-F, PSNR, SSIM, and LPIPS in Table 7. These metrics complement video-level evaluation by assessing frame quality, semantic consistency, and visual changes between the initial and final frames. In particular, PSNR, SSIM, and LPIPS characterize the extent of visual change rather than directly measuring compositional improvement.

Table 6: Quantitative evaluation of the synthetic dataset across five camera-motion categories. Higher scores are better for all metrics.

	VBench 2.0 Metrics						PQA Model Metrics			DAPE
Motion type	Subj. Cons.	Back. Cons.	Mot. Smth.	Dyn. Deg.	Aes. Qual.	Img. Qual.	VQ↑	MQ↑	CA↑	Score↑
Zoom-in	0.9848	0.9912	0.9925	0.9307	0.8030	84.381	0.6110	0.5790	0.5448	1.2011
Shift	0.9783	0.9957	0.9978	0.9411	0.8154	82.548	0.6195	0.5831	0.5380	1.1809
Rotation	0.9742	0.9769	0.9953	0.9331	0.8164	79.563	0.5992	0.5866	0.5266	1.1697
View change	0.9690	0.9588	0.9845	0.9630	0.8033	81.845	0.6007	0.5751	0.5352	1.1736
Compound actions	0.9711	0.9737	0.9806	0.9555	0.7987	81.443	0.6050	0.5633	0.5429	1.1977
Overall	0.9755	0.9793	0.9901	0.9447	0.8074	81.956	0.6071	0.5774	0.5375	1.1846

Table 7: Image-level evaluation of the synthetic dataset across five camera-motion categories. Metric directions follow the main experiments.

Motion type	DeQA↑	Q-Align↑	CLIP-F↑	PSNR↑	SSIM↑	LPIPS↓
Zoom-in	4.104	0.8337	0.9231	13.228	0.4207	0.5148
Shift	4.106	0.8325	0.9222	13.170	0.4189	0.5185
Rotation	4.098	0.8305	0.9217	14.206	0.4366	0.5064
View change	3.997	0.8288	0.9186	12.880	0.4185	0.5190
Compound actions	4.087	0.8302	0.9228	12.866	0.4181	0.5198
Overall	4.078	0.8311	0.9217	13.270	0.4226	0.5157

The results indicate that our synthetic dataset combines favorable visual and aesthetic quality with coherent camera motion and meaningful compositional changes. Together with the human evaluation,

these findings support its suitability for training camera-motion generation models for composition optimization.

C Detailed Motion Trajectory Templates

We employ eight pre-defined camera movement templates as structural priors to guide the initial convergence of the trajectory model. These templates are designed to provide explicit geometric constraints. In the later stages of training, these are transitioned into a unified prompt to promote implicit learning.

Table 8 lists the full text of the eight motion templates used in our experiments.

Table 8: Pre-defined Camera Movement Templates and the Final Unified Prompt.

ID	Template Prompt Text
1	zoom in to change the perspective to find the best composition for the last frame.
2	shift to change the perspective to find the best composition for the last frame.
3	rotate to change the perspective to find the best composition for the last frame.
4	view change to change the perspective to find the best composition for the last frame.
5	rotate and then zoom in to change the perspective to find the best composition for the last frame.
6	shift and then zoom in to change the perspective to find the best composition for the last frame.
7	rotate and then shift to change the perspective to find the best composition for the last frame.
8	view change and then zoom in to change the perspective to find the best composition for the last frame.
Final	**Change the perspective to find the best composition for the last frame.**

C.1 Design Rationale

The construction of these templates follows a structured format: *[Motion Action] + [Goal Suffix]*.

- **Motion Action:** Specifies the camera movement (e.g., *zoom in*, *rotate and then shift*). This helps the model learn the fundamental mapping between text tokens and optical flow patterns.
- **Goal Suffix:** The phrase *"to change the perspective to find the best composition for the last frame"* remains constant across all templates. This provides a consistent semantic context that aligns the motion with the final visual objective.

During the training's transition phase, the **Motion Action** part is gradually masked or replaced, forcing the model to rely on the unified suffix and the visual context of the input frames to infer the optimal trajectory.

D DAPE Model

D.1 Construction of the DAPE Training Dataset

We construct the DAPE training dataset based on the synthetic camera-motion dataset introduced in the main text. For each positive video, we generate three negative counterparts that share the same initial frame but terminate at less favorable compositions. This design controls for the initial scene and composition, providing paired supervision for learning whether a camera trajectory leads to a better compositional outcome.

Zoom-in, shift, and rotation. We randomly select 948 zoom-in videos, 1,360 shift videos, and 1,360 rotation videos as positive examples. For each positive video, we select three alternative endpoint compositions with lower annotated scores from the same source scene in the public composition datasets. Keeping the initial frame fixed, we apply the corresponding video synthesis procedure described in the main text to generate trajectories toward these alternative endpoints. Each positive video is therefore associated with three negative videos depicting the same scene and starting composition, but different, lower-rated final compositions. This yields 3,668 positive videos and 11,004 negative videos across the three categories.

View change. For view-change videos, we follow the keyframe generation and interpolation pipeline described in the main text. We retain the initial frame of each positive video as the shared starting point and use Qwen-Image-Edit to apply perspective rotations to this frame, producing three alternative viewpoints as candidate negative endpoints. These endpoints are intended to retain or exacerbate the compositional limitations of the initial view, rather than reach the favorable endpoint of the positive video. We then use Wan2.2 Team (2025) to interpolate between the shared initial frame and each alternative endpoint, obtaining three negative view-change videos per positive example.

Temporal reversal for direction sensitivity. To encourage direction sensitivity, we additionally reverse each positive video in time and use the reversed sequence as a negative example. The forward and reversed videos contain identical frames but depict opposite compositional transitions: the forward sequence moves toward a favorable composition, whereas its reversal moves away from it. These pairs provide explicit supervision for distinguishing the direction of compositional change while preserving the visual content.

Preference pairs. Each positive video is paired with its three negative counterparts, with the positive video assigned the preferred label. By comparing trajectories from the same starting composition, these pairs provide supervision for distinguishing camera movements that reach favorable compositions from those that terminate at less favorable views.

Table 9: Statistics of the DAPE training dataset. Each positive video is paired with three inferior-endpoint negatives and one temporally reversed negative, yielding four preference pairs.

Motion type	Positive videos	Inferior-endpoint negatives	Reversed negatives	Total videos	Preference pairs
Zoom-in	948	2,844	948	4,740	3,792
Shift	1,360	4,080	1,360	6,800	5,440
Rotation	1,360	4,080	1,360	6,800	5,440
View change	972	2,916	972	4,860	3,888
Total	4,640	13,920	4,640	23,200	18,560

Overall, the dataset comprises 4,640 positive videos and 18,560 negative videos, including 13,920 videos with less favorable endpoint compositions and 4,640 temporally reversed videos, yielding 18,560 preference pairs across four camera-motion categories, as summarized in Table 9.

D.2 Training Details of DAPE

Model architecture. We initialize DAPE from Qwen2-VL-2B-Instruct and attach a bias-free linear reward head to its final hidden representations. Given a video and a composition evaluation instruction, the model produces a scalar reward $r_\theta(v)$ through mean pooling of token-level reward predictions. Higher rewards indicate more favorable composition and camera movement. The evaluation instruction covers spatial layering, geometric harmony, color relationships, frame utilization, visual rhythm, and relationships between scene elements. It additionally emphasizes the direction and magnitude of camera movement, the quality of the resulting composition, and motion smoothness.

Preference learning. We train DAPE using the preference pairs described in Appendix D.1. Each pair consists of a positive video v^+ and a negative counterpart v^-, which either terminates at a less favorable composition or reverses the positive trajectory in time. The training objective encourages the model to assign a higher reward to the positive video. We adopt the Bradley–Terry-with-ties formulation implemented in our training framework. For the strict preference pairs constructed above, the loss is

$$\mathcal{L}_{\mathrm{DAPE}} = -\mathbb{E}_{(v^+,v^-)\sim\mathcal{D}}\left[\log\sigma\left(r_\theta(v^+) - r_\theta(v^-) - \log k\right)\right], \qquad k = 5, \tag{14}$$

where σ denotes the sigmoid function and $\mathcal{D}$ is the training set of preference pairs. The reversed-video pairs explicitly supervise sensitivity to temporal direction, while the alternative-endpoint pairs supervise the quality of the compositional outcome.

Input processing and optimization. We uniformly sample video frames at a target rate of 2 FPS and resize them while preserving their aspect ratio, with a maximum of 200,704 pixels per frame. We adapt the language backbone using LoRA with rank 64, scaling factor 128, and dropout 0.05. The visual encoder is frozen, while the visual merger and reward head are trained jointly with the LoRA parameters. Training runs for 10 epochs on four GPUs using FP16 precision and DeepSpeed ZeRO-2. The batch size is one video pair per GPU with four gradient accumulation steps, resulting in an effective batch size of 16 video pairs. We use a learning rate of 2×10^{-6}, with a linear warm-up over the first 5% of training steps followed by a constant learning rate. Validation and checkpoint saving are performed approximately every 0.3 epoch.

Table 10: Training configuration of DAPE.

Hyperparameter	Value
Backbone	Qwen2-VL-2B-Instruct
Reward dimension	1
Reward pooling	Mean
Frame sampling	Uniform, target 2 FPS
Maximum pixels per frame	200,704
LoRA rank / scaling factor	64 / 128
LoRA dropout	0.05
Visual encoder	Frozen
Visual merger and reward head	Trainable
Preference loss	BTT, $k = 5$
Training epochs	10
Learning rate	2×10^{-6}
Warm-up ratio	0.05
Learning-rate schedule	Constant after warm-up
Number of GPUs	4
Batch size per GPU	1 video pair
Gradient accumulation	4 steps
Effective batch size	16 video pairs
Precision	FP16
Distributed training	DeepSpeed ZeRO-2
Validation / checkpoint interval	Approximately 0.3 epoch

D.3 Direction Discrimination under Temporal Reversal

We evaluate whether DAPE distinguishes compositional improvement from degradation using paired videos with reversed temporal order. Each pair contains exactly the same frames, allowing us to test sensitivity to the direction of compositional change while controlling for visual content.

Independent evaluation set. We construct an evaluation set of 400 video pairs from 400 distinct source scenes excluded from both DAPE and NCG training, with one forward–reverse pair per scene. Candidate trajectories are collected from multiple generation methods to avoid restricting the evaluation to NCG outputs. Three raters independently assess whether the final composition improves over the initial composition. Method identities are concealed, and frame presentation order is randomized. Majority voting determines the label, with disagreements involving three distinct

judgments resolved by an additional rater. We retain trajectories with human-verified compositional improvement and exclude videos containing severe visual artifacts or scene cuts. Selection does not use DAPE scores. For each retained trajectory, we construct an improving sequence and its temporally reversed counterpart:

$$V_i^+ = (x_{i,0}, x_{i,1}, \ldots, x_{i,T}), \qquad V_i^- = (x_{i,T}, x_{i,T-1}, \ldots, x_{i,0}). \tag{15}$$

Thus, V_i^+ progresses from a less favorable composition to a more favorable one, whereas V_i^- reverses this endpoint relationship.

Controlled input processing. We first sample frames using DAPE's evaluation pipeline and then reverse the same sampled frame sequence. Forward and reversed inputs therefore contain identical frames and use the same resolution, frame rate, and encoding settings. All remaining preprocessing and inference settings are identical. DAPE is frozen throughout evaluation, and the evaluation set is not used for model selection or prompt tuning. The two directions are scored independently using the same evaluation prompt, without identifying the forward direction or presenting the paired videos together. Let

$$s_i^+ = \mathrm{DAPE}(V_i^+), \qquad s_i^- = \mathrm{DAPE}(V_i^-), \tag{16}$$

where higher scores indicate a stronger preference for the direction of compositional improvement.

Pairwise direction accuracy. We measure whether DAPE assigns a higher score to the human-verified improving direction:

$$\mathrm{Acc}_{\mathrm{dir}} = \frac{1}{N} \sum_{i=1}^{N} \left[\mathbb{I}(s_i^+ > s_i^-) + \frac{1}{2} \mathbb{I}(s_i^+ = s_i^-) \right], \tag{17}$$

where N is the number of video pairs and $\mathbb{I}(\cdot)$ is the indicator function. Tied scores receive half credit, giving an evaluator that assigns identical scores to both directions an accuracy of 50%. We additionally report the tie rate:

$$R_{\mathrm{tie}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(s_i^+ = s_i^-). \tag{18}$$

Table 11: DAPE direction discrimination on 400 human-verified improving trajectories and their temporal reversals, with one video pair per source scene. Direction accuracy assigns half credit to tied scores.

Evaluator	Scenes	Video pairs	Direction accuracy (%)↑	Tie rate (%)
DAPE	400	400	98.25	0.5

Results and interpretation. As shown in Table 11, DAPE achieves 98.25% direction accuracy with a tie rate of 0.50%. Among the 400 video pairs, it assigns a higher score to the improving direction in 392 cases, a lower score in six cases, and equal scores in two cases. These results demonstrate strong direction discrimination on the evaluated trajectories: despite identical frame content, DAPE generally favors the temporal ordering associated with compositional improvement.

This evaluation supports DAPE's sensitivity to the direction of compositional change on human-verified improving trajectories and their reversals.

E Implementation Details

E.1 Stage 1: SFT

E.1.1 Base Model and Architecture

We fine-tune **HunyuanVideo-1.5**, an 8.3-billion-parameter flow-matching Diffusion Transformer (DiT) that natively supports Image-to-Video (I2V) generation. We use the `480p_i2v_distilled`

checkpoint, a step-distilled variant designed for efficient inference at 480p resolution. The model encodes video through a 3D causal VAE with $16\times$ spatial and $4\times$ temporal compression; text conditioning uses an LLM-based encoder plus a Glyph-byT5-v2 multilingual encoder; image conditioning uses a SigLIP-based vision encoder whose features are injected into the first latent frame.

E.1.2 LoRA Fine-Tuning Configuration

We adapt the transformer via Low-Rank Adaptation (LoRA) applied to *all* linear projection layers (`target_modules="all-linear"`). The rank and scaling factor are set to $r = 64$ and $\alpha = 64$ respectively, yielding an effective LoRA scale of $\alpha/r = 1.0$. No LoRA dropout is applied ($p = 0$) and bias terms are frozen.

E.1.3 Training Data and Resolution

Each sample is an image–prompt pair used in pure I2V mode (`i2v_prob`$= 1.0$). Videos are decoded to $F = 49$ frames (satisfying the VAE's $F = 4n + 1$ requirement) at a resolution of 480×848 pixels (16:9 aspect ratio), which corresponds to the 480p bucket with base size $640^2 \approx 409{,}600$ total pixels.

E.1.4 Flow-Matching Objective

Training follows a continuous-time flow-matching formulation. Clean latents $\mathbf{z}_0$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are linearly interpolated at timestep $t \in (0, 1)$:

$$\mathbf{z}_t = (1 - t)\,\mathbf{z}_0 + t\,\boldsymbol{\epsilon}. \tag{19}$$

The transformer is trained to predict the velocity $\mathbf{v} = \boldsymbol{\epsilon} - \mathbf{z}_0$ under mean-squared error loss:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}}\left[\left\| f_\theta(\mathbf{z}_t, t, \mathbf{c}) - (\boldsymbol{\epsilon} - \mathbf{z}_0)\right\|_2^2\right], \tag{20}$$

where $\mathbf{c}$ denotes the concatenated text, image, and task-conditioning signals. Timesteps are sampled according to a **log-normal** schedule: $u \sim \mathcal{N}(0, 1)$, $t = \sigma(u) = (1 + e^{-u})^{-1}$, which concentrates training mass near intermediate noise levels. Before entering the model, timesteps are further warped by a shift transform

$$t' = \frac{s\,t}{1 + (s - 1)\,t}, \quad s = 3.0, \tag{21}$$

which skews the distribution toward higher noise levels, improving training stability on the step-distilled model.

E.1.5 Optimizer

We use the **Muon** optimizer, a momentum-orthogonalized gradient method. For all weight matrices ($\geq$ 2D parameters), the gradient is first accumulated with Nesterov momentum ($\mu = 0.95$) and then orthogonalized via five Newton–Schulz iterations:

$$\mathbf{U} = \text{NS5}(\hat{\mathbf{G}}), \qquad \theta \leftarrow \theta - \eta_{\text{eff}}\,\mathbf{U} - \eta\,\lambda\,\theta, \tag{22}$$

where $\eta_{\text{eff}} = \eta \cdot 0.2\sqrt{\max(d_{\text{in}}, d_{\text{out}})}$ is a per-layer learning rate that normalizes for weight-matrix size. Scalar and bias parameters ($<$ 2D) are handled by an internal AdamW step with $\beta_1 = \beta_2 = 0.95$, $\epsilon = 10^{-8}$. The learning rate schedule is **constant** after a linear warm-up over 500 steps. Gradients are clipped to a maximum norm of 1.0.

E.1.6 Distributed Training

Training runs on 4×NVIDIA H20 GPUs using **sequence parallelism** (SP) with group size $|\mathcal{G}_{\text{sp}}| = 4$. Under SP, attention tokens are scattered across all four GPUs via all-to-all communication before each attention layer and gathered afterward, so all four GPUs jointly process a single sample. With world size $= 4$ and SP size $= 4$, the effective global batch size is $\lfloor 4/4 \rfloor \times 1 = 1$ sample per step. FSDP is disabled; gradient checkpointing (non-reentrant mode) is applied to every DiT block to reduce activation memory. All model parameters and activations are stored in `bfloat16`; gradient reductions are performed in `float32` for numerical stability. For a more detailed summary of hyperparameters, please refer to Table 12.

Table 12: SFT hyperparameters for HunyuanVideo-1.5 LoRA fine-tuning.

Hyperparameter	**Value**	**Notes**
Base model	`480p_i2v_distilled`	Step-distilled I2V checkpoint
LoRA rank r	64	
LoRA α	64	Scale $= \alpha/r = 1.0$
LoRA target modules	all linear	All projection layers
Resolution	480×848	16:9, 480p
Frames F	49	$F = 4n + 1$
Task mode	I2V only	`i2v_prob`$= 1.0$
Training steps	10,400	Checkpoint every 300 steps
Batch size (per step)	1	SP size = world size
Gradient accumulation	1	
Learning rate η	1×10^{-4}	Constant after warm-up
Warm-up steps	500	Linear
Weight decay λ	0.01	
Max grad norm	1.0	
Optimizer	Muon	Nesterov $\mu = 0.95$, NS steps$= 5$
Timestep sampling	Log-normal	$u \sim \mathcal{N}(0, 1)$, $t = \sigma(u)$
Timestep shift s	3.0	Training; 5.0 for inference
Sequence parallelism	4 GPUs	All-to-all in attention
FSDP	Disabled	
Gradient checkpointing	Enabled	Per DiT block, non-reentrant
Precision	BF16	Reduce in FP32

E.2 Stage 2: GRPO

E.2.1 Base Model

We fine-tune HunyuanVideo-1.5 in its image-to-video (I2V) distilled LoRA variant. All training videos are generated at a resolution of 480×864 with $F = 49$ frames ($F = 4n + 1$, $n = 12$) at 24 fps. The base model uses classifier-free guidance scale $w = 1.0$ (distilled) and flow-shift $\tau = 5.0$.

E.2.2 GRPO Rollout

We employ SAGE-GRPO (`sde_type=sage_grpo`) as the policy-gradient estimator. Each rollout performs $T_{\text{roll}} = 20$ denoising steps using the flow-matching discrete scheduler with flow-shift 5.0. Timesteps are sampled according to a log-normal distribution. For each training prompt, $G = 4$ videos are generated per step (group size), and advantages are computed within the group via mean-centering and standard-deviation normalization. Advantages are clipped to $[-A_{\max}, A_{\max}]$ with $A_{\max} = 5.0$. The importance-weight ratio is clipped within $[1 - \epsilon, 1 + \epsilon]$, $\epsilon = 10^{-4}$. Gradient balancing across timesteps is enabled, and timestep permutation augmentation is applied during rollouts.

E.2.3 LoRA Configuration

We apply Low-Rank Adaptation (LoRA) to all linear layers of the transformer, using rank $r = 64$, scaling factor $\alpha = 64$, and dropout $p = 0$.

E.2.4 Reward Model

Videos are scored by `PQA` (`videoalign_ca_local`, a VLM-based reward model that produces three normalized dimension scores: Visual Quality (VQ), Motion Quality (MQ), and Composition Aesthetic (CA). The composite reward is:

$$r = w_{\text{VQ}}\,\hat{r}_{\text{VQ}} + w_{\text{MQ}}\,\hat{r}_{\text{MQ}} + w_{\text{CA}}\,\hat{r}_{\text{CA}}, \quad w_{\text{VQ}} = 1.0,\ w_{\text{MQ}} = 1.0,\ w_{\text{CA}} = 2.0, \tag{23}$$

where $\hat{r}_\cdot$ denotes per-dimension z-score normalization against model-specific statistics.

E.2.5 KL REGULARIZATION

To prevent reward hacking while preserving visual quality, we apply a dual KL penalty:

$$\mathcal{L}_{\mathrm{KL}} = \beta\Big[\lambda_{\mathrm{mv}}\,\mathrm{KL}_{\mathrm{moving}} + \lambda_{\mathrm{step}}\,\mathrm{KL}_{\mathrm{step}}\Big], \tag{24}$$

with base weight $\beta = 10^{-5}$, KL coefficient $c = 10^{-7}$, moving-average weight $\lambda_{\mathrm{mv}} = 1.0$, and per-step weight $\lambda_{\mathrm{step}} = 0.1$. The KL is computed against a reference model updated every $\Delta_{\mathrm{ref}} = 200$ steps, with an adaptive PID controller.

E.2.6 OPTIMIZER AND TRAINING SCHEDULE

We use the Muon optimizer with learning rate $\eta = 5 \times 10^{-5}$, weight decay $\lambda = 0.01$, and gradient clipping at norm 1.0. total training lasts 150 steps with batch size of 1.

E.2.7 TRAINING INFRASTRUCTURE

Experiments are conducted on 4 NVIDIA H20 GPUs. The transformer is sharded with FSDP2, applied block-by-block to `double_blocks` and `single_blocks`, under a mixed-precision policy of `bfloat16` parameters and `float32` gradient reductions. Gradient checkpointing is enabled to reduce activation memory. The reference model is CPU-offloaded during policy updates. Sequence parallelism is disabled (`sp_size` $= 1$), and all experiments use random seed 42.

F COMPREHENSIVE QUANTITATIVE RESULTS AND HUMAN STUDY

This section provides quantitative results for all 11 methods, complementing the selected comparisons in the main text, followed by a human study of composition optimization quality.

F.1 COMPREHENSIVE QUANTITATIVE RESULTS

Due to space limitations, the main text presents comparisons with a subset of methods. Here, we provide the complete quantitative comparison across 11 methods: PPC Yao et al. (2025), Seedance2.0, HappyHorse, S2CNet (Su et al., 2024), Venus (Du et al., 2026), ProCrop Zhang et al. (2026), UNIC Liu et al. (2023), CACNet Hong et al. (2021a), CGS Li et al. (2020a), GAIC Zeng et al. (2022), and our NCG. All methods are evaluated on the same input images, following the evaluation protocol described in the main text.

Table 13 summarizes the results in two parts. The upper part evaluates the four video-based methods using VBench 2.0, PQA, and DAPE, covering video quality, temporal consistency, and the direction of compositional improvement. The lower part reports DeQA, Q-Align, CLIP-F, PSNR, SSIM, and LPIPS for all 11 methods. For video-based methods, the final frame serves as the optimized composition; for image-based methods, we use the output image. PSNR, SSIM, and LPIPS quantify visual changes relative to the input rather than directly measuring aesthetic improvement.

Among the evaluated video-based methods, NCG achieves the highest PQA scores across all three dimensions and the highest DAPE score of 0.9328. These results support its effectiveness in producing visually favorable videos with camera trajectories that improve composition.

F.2 HUMAN STUDY

Participants and evaluation samples. We conduct a human study to assess the perceived quality of composition optimization. 16 participants take part in the study: 5 computer vision experts, 3 photography experts, and 8 general users. We randomly select 18 input images and apply 11 methods to each image: PPC, Seedance, HappyHorse, S2CNet, Venus, ProCrop, UNIC, CACNet, CGS, GAIC, and our NCG. This produces 198 results in total, organized into 18 groups, each containing the outputs of all 11 methods for the same input image.

Ranking protocol. Each participant independently ranks the 11 results within each group according to their overall compositional quality, from the most preferred to the least preferred, without ties.

Table 13: Comprehensive performance comparison of composition optimization. The upper part reports video-level metrics for video-based methods, while the lower part reports image-level metrics for all 11 methods.

	VBench2.0 Metrics						PQA Model Metrics			DAPE Metric
Model	Subj. Cons.	Back. Cons.	Mot. Smth.	Dyn. Deg.	Aes. Qual.	Img. Qual.	VQ↑	MQ↑	CA↑	DAPE↑
PPC	0.8870	0.9235	0.9887	0.2403	0.5381	63.069	0.3872	0.3677	0.3680	0.8631
Seedance2.0	0.9292	0.9510	0.9901	0.8214	0.5507	64.084	0.3610	0.3472	0.3618	0.7816
HappyHorse	0.8668	0.9186	0.9799	0.9697	0.5406	63.495	0.3500	0.3353	0.3426	0.7380
NCG (Ours)	0.9491	0.9530	0.9892	0.8036	0.5535	63.388	0.4947	0.4941	0.4828	0.9328

Model	DeQA↑	Q-Align↑	CLIP-F↑	PSNR↑	SSIM↑	LPIPS↓
PPC	3.476	0.7226	0.8589	14.137	0.4378	0.5019
Seedance2.0	**3.861**	**0.8163**	<u>0.8907</u>	13.068	0.4269	0.5326
HappyHorse	3.381	0.7332	0.7861	10.720	0.4154	0.6896
S2CNet	3.352	0.7118	0.8868	15.035	0.4379	0.4167
Venus	3.364	0.7093	0.8676	13.160	0.4533	0.4970
ProCrop	3.417	0.7207	0.8713	14.360	0.4768	0.4570
UNIC	3.265	0.7044	0.8478	<u>15.058</u>	<u>0.4891</u>	<u>0.3953</u>
CACNet	3.208	0.6862	0.8065	**15.154**	**0.4940**	**0.3933**
CGS	3.188	0.6822	0.8033	14.623	0.4723	0.4195
GAIC	3.054	0.6709	0.7741	14.181	0.4707	0.4439
NCG (Ours)	<u>3.561</u>	<u>0.7374</u>	**0.9020**	13.459	0.4270	0.5092

Comparing results derived from the same input image allows participants to judge the relative effectiveness of different methods under a shared scene context. Each participant completes all 18 groups, yielding 288 complete rankings across the study.

Score aggregation. We convert each ranking into a numerical preference score: the first-ranked result receives 11 points, the second receives 10 points, and the last receives 1 point. Specifically, let $r_{p,i,m} \in \{1, \ldots, 11\}$ denote the rank assigned by participant p to method m on image i. The overall score for method m is

$$S_m = \frac{1}{16 \times 18} \sum_{p=1}^{16} \sum_{i=1}^{18} (12 - r_{p,i,m}) \,. \tag{25}$$

Thus, each method receives 288 individual scores, and its final score ranges from 1 to 11, with higher values indicating stronger human preference. We additionally report scores averaged within each participant group to characterize preferences across different backgrounds. The results are summarized in Table 14.

G Camera Trajectory OOD Evaluation: Methodology and Metrics

This appendix provides a detailed account of the Out-of-Distribution (OOD) evaluation pipeline used to assess camera motion quality in generated videos, as referenced in the main text (Tab. 3 and Fig. 6).

G.1 Overview

To evaluate whether a generated video faithfully executes a prescribed camera motion, we design a four-stage pipeline: (1) 3D camera pose estimation via DUSt3R, (2) pose sequence feature extraction via a 1D-ResNet-18 classifier, (3) energy-based OOD scoring, and (4) metric computation against a reference distribution derived from real training data. The core insight is that a model trained to discriminate eight canonical camera-motion categories induces a feature space in which in-distribution

Table 14: Human evaluation of composition optimization. Scores are derived from rankings of 11 methods and averaged across the 18 input images and the corresponding participants. Higher scores indicate stronger preference.

Method	Computer vision experts	Photography experts	General users	Overall↑
PPC	6.333	6.000	6.208	6.208
Seedance	5.078	5.833	6.090	5.726
HappyHorse	4.811	2.167	5.292	4.556
S2CNet	6.367	6.278	6.299	6.316
Venus	5.867	6.056	5.986	5.962
ProCrop	5.689	5.389	5.569	5.573
UNIC	5.778	5.056	6.278	5.892
CACNet	5.722	6.056	5.757	5.802
CGS	5.644	6.167	5.764	5.802
GAIC	6.056	6.111	5.333	5.705
NCG (Ours)	8.656	10.889	7.424	8.458

trajectories cluster tightly, while anomalous or malformed trajectories exhibit high *energy* and reside in low-density peripheral regions.

G.2 Stage 1: Camera Pose Estimation

Given a generated video, we uniformly sample 49 frames and feed them into DUSt3R Wang et al. (2024a), a feed-forward stereo reconstruction network that jointly estimates per-frame camera poses in a common world coordinate system. DUSt3R processes all pairwise frame combinations via a complete scene graph and refines the poses through a global alignment step (PointCloudOptimizer, 300 iterations, cosine learning-rate schedule). Each frame t yields a camera-to-world transformation matrix $\mathbf{T}_t \in \mathrm{SE}(3)$, from which we extract the translation vector $\mathbf{p}_t \in \mathbb{R}^3$ and rotation quaternion $\mathbf{q}_t \in \mathbb{R}^4$ (converted from the rotation matrix via the Shepperd method). The raw pose sequence is represented as $\mathbf{X} = \{[\mathbf{p}_t; \mathbf{q}_t]\}_{t=1}^{N} \in \mathbb{R}^{N\times 7}$.

G.3 Stage 2: Pose Sequence Preprocessing and Feature Extraction

Normalization. To achieve translation invariance and scale invariance, we first subtract the initial position $\mathbf{p}_1$ from all positions, then normalize by the end-to-end trajectory length:

$$\tilde{\mathbf{p}}_t = \frac{\mathbf{p}_t - \mathbf{p}_1}{\|\mathbf{p}_N - \mathbf{p}_1\|_2 + \epsilon}. \tag{26}$$

Resampling. All sequences are resampled to a fixed length of $L = 45$ via linear interpolation along the temporal axis, yielding $\tilde{\mathbf{X}} \in \mathbb{R}^{7\times 45}$.

Classifier. We train a 1D-ResNet-18 with 7-channel input on a dataset of real camera trajectories labelled with eight canonical motion categories: *zoom in*, *rotate*, *shift*, *view change*, *rotate-then-zoom-in*, *rotate-then-shift*, *shift-then-zoom-in*, and *view-change-then-zoom-in*. The network maps each pose sequence to a logit vector $\mathbf{f}(\tilde{\mathbf{X}}) \in \mathbb{R}^8$ and a 512-dimensional penultimate feature vector used for visualization (Fig. 6).

G.4 Stage 3: Energy-Based OOD Scoring

We use an energy-based criterion Liu et al. (2020) to assess whether generated camera trajectories fall within the reference distribution represented by the trained motion classifier. For a pose sequence with logits $\mathbf{z} = \mathbf{f}(\tilde{\mathbf{X}})$, the energy score is

$$E(\mathbf{z}) = -T \log \sum_{c=1}^{C} \exp(z_c/T), \tag{27}$$

where $C = 8$ is the number of motion categories and $T = 1$ is the temperature. Under this scoring convention, higher energy is treated as stronger evidence of distributional mismatch. Energy depends on the absolute logit values and is not equivalent to softmax entropy or maximum softmax probability.

Threshold calibration. We calibrate a single threshold τ using the 95th percentile of energy scores on the held-out validation split of the reference trajectory dataset:

$$\tau = \text{Percentile}_{95}\left(\{E(\mathbf{z}_i)\}_{i \in \mathcal{D}_{\text{val}}}\right). \tag{28}$$

The classifier and threshold are fixed throughout evaluation. A generated video is classified as out-of-distribution if $E(\mathbf{z}) > \tau$; otherwise, it is accepted by the detector. This calibration targets an empirical rejection rate of approximately 5% on the reference validation split, subject to finite-sample quantile estimation and tied scores.

G.5 Evaluation Metrics

We report four metrics for each evaluated set of N generated videos. ORR and LR summarize the same energy-based detection decisions and are exact complements. AC describes classifier confidence, while MS summarizes normalized frame-to-frame translation displacement.

OOD Rejection Rate (ORR). ORR is the percentage of generated videos whose trajectory energy exceeds the fixed threshold τ defined in Eq. equation 28:

$$\text{ORR} = \frac{|\{i : E(\mathbf{z}_i) > \tau\}|}{N} \times 100\%. \tag{29}$$

A lower ORR indicates that fewer generated trajectories are rejected by the energy-based detector.

Legal Rate (LR). We use the term Legal Rate to denote the percentage of generated videos accepted by the same detector:

$$\text{LR} = \frac{|\{i : E(\mathbf{z}_i) \leq \tau\}|}{N} \times 100\% = 100\% - \text{ORR}. \tag{30}$$

LR does not require ground-truth motion labels. A higher LR indicates that a larger proportion of generated trajectories fall within the detector's energy acceptance region.

Average Certainty (AC). AC is the mean maximum softmax probability across all samples:

$$\text{AC} = \frac{1}{N} \sum_{i=1}^{N} \max_c \text{softmax}(\mathbf{z}_i)_c. \tag{31}$$

AC describes how confidently the classifier assigns generated trajectories to the predefined motion categories. For $C = 8$, a uniform class distribution gives $\text{AC} = 1/8 = 0.125$.

G.6 Motion Smoothness

We evaluate Motion Smoothness (MS) along three components: camera translation, camera rotation, and zoom. For each component, we measure changes in acceleration using a dimensionless jerk score. The three scores are reported separately as MS_{trans}, MS_{rot}, and MS_{zoom}. When all three components are active and reliably estimated, the overall score is their arithmetic mean:

$$\text{MS} = \frac{\text{MS}_{\text{trans}} + \text{MS}_{\text{rot}} + \text{MS}_{\text{zoom}}}{3}. \tag{32}$$

Lower values indicate smaller normalized jerk. This aggregate assigns equal weights to the three components.

Trajectory representation. Consider a video sampled at $L \geq 4$ equally spaced timestamps $t_k = k\Delta t$, where $k = 0, \ldots, L-1$ and $T = (L-1)\Delta t$ is the evaluated duration. At each timestamp, let $\mathbf{p}_k \in \mathbb{R}^3$ denote the camera position, $\mathbf{R}_k \in \text{SO}(3)$ the camera-to-world rotation, and $f_k > 0$ the focal length expressed under a consistent image-coordinate convention. We represent zoom using the log-focal-length signal

$$z_k = \log\left(\frac{f_k}{f_0}\right). \tag{33}$$

Using log focal length expresses zoom in terms of relative rather than absolute focal-length changes.

Translational motion. The translational velocity is

$$\mathbf{v}_k = \frac{\mathbf{p}_{k+1} - \mathbf{p}_k}{\Delta t}, \qquad k = 0, \dots, L-2. \tag{34}$$

We estimate translational jerk by the second finite difference of velocity:

$$\mathbf{j}_k^{\text{trans}} = \frac{\mathbf{v}_{k+2} - 2\mathbf{v}_{k+1} + \mathbf{v}_k}{\Delta t^2}, \qquad k = 0, \dots, L-4. \tag{35}$$

The translational motion amplitude is defined as the accumulated path length:

$$S_{\text{trans}} = \sum_{k=0}^{L-2} \|\mathbf{p}_{k+1} - \mathbf{p}_k\|_2 . \tag{36}$$

For non-negligible translational motion, the score is

$$\text{MS}_{\text{trans}} = \frac{T^5}{S_{\text{trans}}^2} \Delta t \sum_{k=0}^{L-4} \left\|\mathbf{j}_k^{\text{trans}}\right\|_2^2 . \tag{37}$$

We normalize by accumulated path length rather than endpoint displacement so that a trajectory returning near its starting position does not become singular solely because its endpoints are close.

Rotational motion. We compute relative rotations in the world coordinate frame:

$$\Delta\mathbf{R}_k = \mathbf{R}_{k+1}\mathbf{R}_k^\top . \tag{38}$$

The angular velocity is

$$\boldsymbol{\omega}_k = \frac{\text{Log}\left(\Delta\mathbf{R}_k\right)^\vee}{\Delta t}, \qquad k = 0, \dots, L-2, \tag{39}$$

where Log is the rotation-matrix logarithm and $(\cdot)^\vee$ maps a skew-symmetric matrix to its corresponding three-dimensional rotation vector. The sampling rate must be sufficient for consecutive rotations to remain below the π ambiguity of the principal logarithm.

Angular jerk is estimated as

$$\mathbf{j}_k^{\text{rot}} = \frac{\boldsymbol{\omega}_{k+2} - 2\boldsymbol{\omega}_{k+1} + \boldsymbol{\omega}_k}{\Delta t^2}, \qquad k = 0, \dots, L-4. \tag{40}$$

Using the accumulated rotation angle

$$S_{\text{rot}} = \sum_{k=0}^{L-2} \left\|\text{Log}\left(\Delta\mathbf{R}_k\right)^\vee\right\|_2 , \tag{41}$$

we define

$$\text{MS}_{\text{rot}} = \frac{T^5}{S_{\text{rot}}^2} \Delta t \sum_{k=0}^{L-4} \left\|\mathbf{j}_k^{\text{rot}}\right\|_2^2 . \tag{42}$$

This formulation avoids discontinuities caused by Euler-angle wrapping or quaternion sign changes.

Zoom motion. The zoom velocity and jerk are

$$u_k = \frac{z_{k+1} - z_k}{\Delta t}, \qquad k = 0, \dots, L-2, \tag{43}$$

$$j_k^{\text{zoom}} = \frac{u_{k+2} - 2u_{k+1} + u_k}{\Delta t^2}, \qquad k = 0, \dots, L-4. \tag{44}$$

The zoom amplitude is the total variation of log focal length:

$$S_{\text{zoom}} = \sum_{k=0}^{L-2} |z_{k+1} - z_k|. \tag{45}$$

We define

$$\text{MS}_{\text{zoom}} = \frac{T^5}{S_{\text{zoom}}^2} \Delta t \sum_{k=0}^{L-4} |j_k^{\text{zoom}}|^2 . \tag{46}$$

Zoom smoothness requires a reliable focal-length signal. If only image-based scale changes are available, the resulting score is reported as apparent zoom smoothness, since image scale can also change through camera translation and scene geometry.

Normalization and inactive components. The factor T^5/S_d^2, where $d \in \{\mathrm{trans}, \mathrm{rot}, \mathrm{zoom}\}$, makes each score dimensionless. In the continuous-time limit, this normalization removes the effect of uniform temporal scaling and constant amplitude scaling of the corresponding motion signal. It does not make scores independent of trajectory shape.

We compute a component score only when its motion amplitude is reliably distinguishable from estimation noise. The activity criterion for each component is calibrated on independent validation data and fixed across methods. For translation, this criterion must account for the scale ambiguity of reconstructed trajectories. Inactive components are marked as not applicable rather than assigned a zero score.

G.7 VISUALIZATION

To visualize generated trajectories, we extract the 512-dimensional penultimate features from the trained ResNet-18 for all test samples and compute a two-dimensional t-SNE embedding (perplexity= 30, 2000 iterations, PCA initialization). In Fig. 6, the horizontal coordinate is the first t-SNE coordinate, while the vertical coordinate is the raw energy score $E(\mathbf{z})$.

The dashed horizontal line marks the fixed threshold τ defined in Eq. equation 28. Samples above this line satisfy $E(\mathbf{z}) > \tau$ and are classified as OOD. The t-SNE coordinate is used only for visualization and does not affect OOD classification.

H ABLATION STUDY

H.1 ANALYSIS OF TRAINING STAGE.

We conducted comprehensive ablation studies for each stage of our model, with the quantitative results reported in Table 15. The data demonstrate that our final model consistently outperforms all intermediate variants across aesthetic metrics. Notably, the most significant performance gains are observed in camera trajectory consistency and rationality under Out-of-Distribution (OOD) scenarios, highlighting the effectiveness of our proposed guidance mechanism.

Table 15: Performance comparison of composition optimization.

	VBench2.0 Metrics						PQA Model Metrics		
Model	Subj. Cons.	Back. Cons.	Mot. Smth.	Dyn. Deg.	Aes. Qual.	Img. Qual.	**VQ↑**	**MQ↑**	**CA↑**
w/o SFT	0.9481	0.9468	0.9894	0.7251	0.5506	62.840	0.4456	0.4515	0.4776
w/o GRPO	0.9452	0.9473	**0.9913**	**0.8184**	0.5510	62.546	0.3946	0.3905	0.4349
NCG (Ours)	**0.9491**	**0.9530**	0.9892	0.8036	**0.5535**	**63.388**	**0.4947**	**0.4941**	**0.4828**

Model	DeQA↑	Q-Align↑	CLIP-F↑	PSNR↑	SSIM↑	LPIPS↓
w/o SFT	3.482	0.7186	0.8978	14.779	0.4464	0.4697
w/o GRPO	3.037	0.7090	0.8968	13.721	0.4394	0.5100
NCG (Ours)	**3.561**	**0.7374**	**0.9020**	13.459	0.4270	0.5092

H.2 ANALYSIS OF GRADIENT INTERFERENCE.

We conducted a quantitative evaluation of gradient interference, as summarized in Table 16. Our empirical findings reveal that the model achieves peak performance exclusively at a batch size of 1, with a significant degradation in aesthetics and visual perception metrics as the batch size increases. This phenomenon stems from the inherent non-convexity of the one-to-many mapping problem. Specifically, our dataset contains multiple video trajectories originating from the same initial frame but converging toward distinct local optima (i.e., different aesthetic compositions). Under a stochastic batch loading regime, samples with identical starting points may be grouped within the same batch, triggering contradictory gradient directions. This gradient interference leads to optimization 'ambiguity,' which manifests as severe visual artifacts, including temporal stretching, geometric distortion, and blurring.

Table 16: Performance comparison of composition optimization.

Model	VBench2.0 Metrics						PQA Model Metrics		
	Subj. Cons.	Back. Cons.	Mot. Smth.	Dyn. Deg.	Aes. Qual.	Img. Qual.	**VQ↑**	**MQ↑**	**CA↑**
BatchSize=4	0.8843	0.9207	0.9863	0.6958	0.5126	60.090	0.2918	0.2964	0.3255
BatchSize=2	0.8977	0.9265	0.9891	0.7079	0.5138	60.979	0.3292	0.3251	0.3401
BatchSize=1	**0.9491**	**0.9530**	**0.9892**	**0.8036**	**0.5535**	**63.388**	**0.4947**	**0.4941**	**0.4828**

Model	DeQA↑	Q-Align↑	CLIP-F↑	PSNR↑	SSIM↑	LPIPS↓
BatchSize=4	3.208	0.6338	0.8437	13.508	0.4351	0.5413
BatchSize=2	3.217	0.6406	0.8483	13.779	0.4372	0.5356
BatchSize=1	3.561	0.7374	0.9020	13.459	0.4270	0.5092

H.3 Aesthetic Convergence Analysis

H.3.1 Terminal Motion Settling

Conceptual interpretation. The aesthetic manifold $\mathcal{M}_A$ provides a conceptual description of favorable compositions and the desired navigation objective. It is not an explicitly parameterized or separately learned component of our architecture. In practice, supervised fine-tuning and reward-based preference optimization encourage the model to generate trajectories toward favorable compositions using computable aesthetic and quality signals. We neither explicitly recover the geometry of $\mathcal{M}_A$ nor compute the gradient of the abstract aesthetic function A. Accordingly, the following analysis examines whether motion decreases near the end of a generated trajectory, providing an operational assessment of terminal settling.

Evaluation protocol and metrics. We compare PPC, Seedance, HappyHorse, and NCG using optical-flow statistics over the complete video and its final eight frames. Optical-flow magnitude measures apparent image motion between adjacent frames and serves as a proxy for motion intensity. For each video, we compute its mean motion intensity over the full trajectory and the terminal segment. We then average these per-video measurements across the evaluated videos to obtain M_{full} and M_{end}, respectively. We report the following metrics:

- **Full-trajectory intensity** (M_{full}) characterizes the average motion intensity over the complete trajectory. It provides a reference scale for interpreting terminal motion, with no inherently preferred direction.
- **Terminal intensity** (M_{end}) characterizes the average residual motion over the final eight frames. Lower values indicate less movement near the trajectory endpoint.
- **Intensity decay** measures the absolute decrease from full-trajectory to terminal motion intensity:
$$D_{\text{motion}} = M_{\text{full}} - M_{\text{end}}. \tag{47}$$
The percentage reported in parentheses is the corresponding relative reduction:
$$\Delta_{\text{motion}} = \frac{M_{\text{full}} - M_{\text{end}}}{M_{\text{full}}} \times 100\%. \tag{48}$$
Larger positive values indicate a stronger decrease in motion toward the endpoint.
- **Settling ratio** expresses terminal motion intensity relative to the full-trajectory reference:
$$R_{\text{settle}} = \frac{M_{\text{end}}}{M_{\text{full}}}. \tag{49}$$
A ratio below one indicates reduced terminal motion, while a ratio close to one indicates little change. Lower ratios indicate stronger relative settling. This ratio is computed from the aggregate mean intensities rather than by averaging per-video ratios.
- **Across-video terminal variance** measures the variance of per-video mean terminal motion intensity across the evaluated videos. It describes how residual motion differs between

inputs, rather than temporal fluctuations within an individual video. We report this metric as a descriptive statistic without assigning a preferred direction: a smaller value indicates more concentrated terminal intensities, whereas a larger value indicates greater variation across videos.

Table 17: Optical-flow analysis of terminal settling across four video-based methods. Terminal statistics use the final eight frames of each video. Percentages in parentheses indicate motion reduction relative to the full-trajectory intensity. Across-video terminal variance measures variation between per-video terminal means.

Metric	PPC	Seedance	HappyHorse	**NCG (Ours)**
Full-trajectory intensity	2.5707	3.4270	3.1065	2.9283
Terminal intensity↓	2.5697	2.9757	2.6592	**1.9098**
Intensity decay↑	0.0010 (0.04%)	0.4513 (13.17%)	0.4473 (14.40%)	**1.0185 (34.78%)**
Settling ratio↓	0.9996	0.8683	0.8560	**0.6522**
Across-video terminal variance	3.5325	3.1648	2.7694	2.4254

Results and interpretation. Table 17 compares terminal settling across the four methods. NCG reduces its mean motion intensity from 2.9283 over the full trajectory to 1.9098 near the endpoint, corresponding to an absolute decrease of 1.0185 and a relative reduction of 34.78%. In comparison, Seedance and HappyHorse exhibit reductions of 13.17% and 14.40%, respectively, while PPC remains nearly unchanged with a reduction of only 0.04%. NCG achieves both the lowest terminal intensity and the lowest settling ratio of 0.6522, indicating a stronger relative decrease in motion toward the trajectory endpoint.

NCG also exhibits the lowest across-video terminal variance of 2.4254, indicating that its per-video terminal motion intensities are more concentrated across the evaluated inputs. This statistic characterizes differences between videos; it does not directly measure temporal smoothness or establish that motion intensity is appropriately adapted to scene content.

Scope and limitations. The terminal motion intensity remains nonzero, and the reported aggregate statistics do not imply that every trajectory comes to a complete stop. The results support finite-horizon terminal settling rather than formal asymptotic convergence. Furthermore, reduced optical flow alone does not establish aesthetic optimality, since low motion can also occur in videos without meaningful compositional improvement. This analysis therefore complements the aesthetic evaluations in the main text, without constituting a proof of convergence to an aesthetic optimum or explicit recovery of the aesthetic manifold.

H.3.2 Motion Adaptation to Initial Composition Quality

We examine whether camera motion adapts to initial composition quality using paired inputs from the same scene. Complementing terminal settling analysis, this experiment tests whether the model limits unnecessary motion on favorable inputs while effectively improving perturbed compositions.

Paired-input construction. We select 100 photographs with favorable compositions, excluded from all training data and test sets used in other experiments, including those for NCG and DAPE. Each photograph serves as a favorable input x_i^{good}, from which we construct a perturbed input x_i^{pert} through controlled changes in framing, translation, or rotation, yielding 100 scene pairs. Both inputs are standardized to the same resolution and aspect ratio. Human assessment verifies that the perturbation reduces compositional quality without introducing severe image artifacts. The favorable input is used to construct the pair, rather than treated as a unique target composition.

Generation protocol. We compare PPC, Seedance, HappyHorse, NCG without GRPO, and the complete NCG model. Within each method, both input conditions receive the same composition-improvement instruction, without specifying whether the initial composition is favorable or perturbed. Each method generates one video per input, yielding 200 videos per method. Video duration, evaluation frame rate, and spatial resolution are standardized across methods. All generated samples are included in the evaluation.

Motion adaptation. We reuse the full-trajectory motion intensity M_{full} defined above. For each scene, we compute the full-trajectory motion intensities m_i^{good} and m_i^{pert} for the two input conditions. We then compute the paired difference:

$$\Delta M_i = m_i^{\text{pert}} - m_i^{\text{good}}, \qquad \overline{\Delta M} = \frac{1}{N}\sum_{i=1}^{N} \Delta M_i, \tag{50}$$

where $N = 100$ is the number of scene pairs. A positive $\overline{\Delta M}$ indicates greater motion for perturbed inputs than for favorable inputs. We report a 95% confidence interval obtained by bootstrapping scene pairs, retaining all samples from each scene together.

We do not use the terminal-to-full motion ratio as the primary measure in this experiment. For an already favorable input, consistently low motion can produce a ratio close to one despite appropriate behavior. Furthermore, a larger $\overline{\Delta M}$ is not inherently better, since excessive movement on perturbed inputs can also increase this difference.

Compositional preservation and improvement. Three raters independently compare the initial and final frames of each generated video. The frames are presented in randomized left–right order, with method identities and input conditions concealed. Raters select the better composition or indicate a tie, focusing on subject placement, visual balance, framing, and spatial relationships rather than image sharpness or rendering quality. Responses are mapped to improved, approximately unchanged, or degraded relative to the initial frame. The majority vote determines the final label; cases in which all three raters select different options are adjudicated by an additional rater. Evaluation compares each output with its own input, without requiring recovery of the favorable reference image.

For favorable inputs, the non-degradation rate is the proportion of outputs labeled improved or approximately unchanged. For perturbed inputs, the improvement rate is the proportion labeled improved.

Table 18: Motion adaptation to initial composition quality on 100 held-out scene pairs. Good and Perturbed denote favorable and perturbed inputs from the same scene. Motion columns report mean full-trajectory intensity and the paired difference as mean [95% CI], with confidence intervals obtained by resampling scene pairs. Human evaluation reports non-degradation on favorable inputs and improvement on perturbed inputs.

	Motion intensity			Human evaluation	
Method	Good M_{full}	Perturbed M_{full}	Paired $\overline{\Delta M}$ (95% CI)	Good: non-degradation rate (%)↑	Perturbed: improvement rate (%)↑
PPC	2.5811	2.5638	-0.0173 [-0.0751, -0.0004]	53%	63%
Seedance	3.5612	3.4930	-0.0682 [-0.1115, -0.0475]	41%	46%
HappyHorse	3.2172	3.1752	-0.0420 [-0.0901, -0.0309]	36%	33%
NCG w/o GRPO	2.4421	2.7161	0.2740 [0.2491, 0.2902]	55%	51%
NCG (Ours)	1.8305	2.7906	0.9601 [0.9397, 0.9732]	73%	82%

Results and interpretation. Table 18 shows that NCG produces less motion on favorable inputs ($M_{\text{full}} = 1.8305$) than on perturbed inputs (2.7906), with a paired difference of $\overline{\Delta M} = 0.9601$. NCG without GRPO exhibits a smaller difference of 0.2740, whereas PPC, Seedance, and HappyHorse show small negative differences, indicating no corresponding increase in mean motion for perturbed inputs.

Human evaluation complements these motion statistics. NCG achieves the highest observed non-degradation rate on favorable inputs (73%) and improvement rate on perturbed inputs (82%). Compared with NCG without GRPO, these rates increase by 18 and 31 percentage points, respectively. Thus, the larger motion difference is accompanied by better compositional outcomes in both input conditions, supporting the contribution of GRPO to content-adaptive adjustment.

These results support adaptation of motion intensity to initial composition quality rather than indiscriminate motion suppression. However, favorable compositions still degrade in 27% of NCG outputs, and 18% of perturbed inputs do not improve, indicating that this behavior is not reliable for every input.

Joint analysis. Taken together, the main-text aesthetic evaluations and the two analyses above provide complementary evidence for content-adaptive compositional guidance. The main-text results support the aesthetic quality and direction of compositional adjustment. Terminal motion analysis (Table 17) shows a 34.78% reduction in NCG's motion intensity near the endpoint, indicating that generated trajectories tend to settle within the evaluated time horizon. The paired-input experiment (Table 18) further shows that NCG uses less motion for favorable initial compositions, while preserving their quality in 73% of cases and improving perturbed compositions in 82% of cases. These findings jointly support three desirable behaviors: compositional improvement, terminal settling, and adaptation of motion intensity to the need for adjustment. They therefore provide empirical support for content-adaptive convergence in an operational, finite-horizon sense. Because these evaluations summarize complementary outcomes across samples, they do not establish that every trajectory improves monotonically or reaches a local aesthetic optimum.