AES6D: A DATASET AND BENCHMARK FOR AESTHETIC-GUIDED 6-DOF IMAGE EDITING

Anonymous authors
Paper under double-blind review

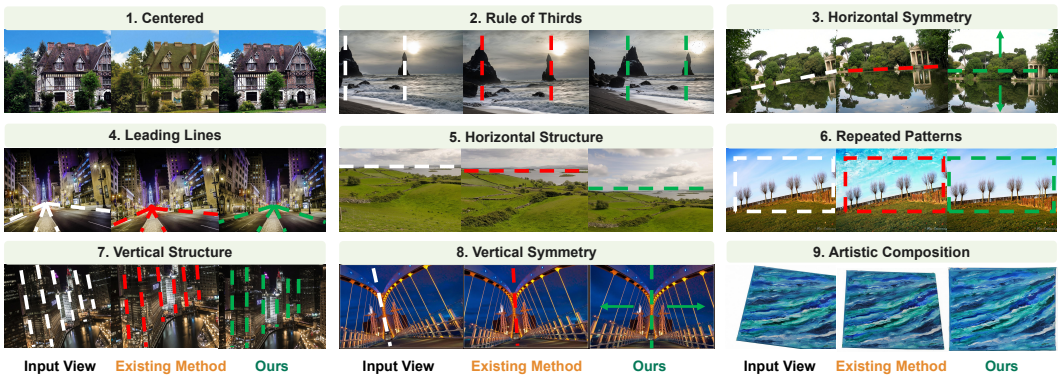


Figure 1: Results across nine composition patterns in Aes6D. Models trained on Aes6D perform fine-grained 6-DoF viewpoint adjustments for different composition issues, achieving more substantial composition improvement than existing methods while better preserving the original scene content.

ABSTRACT

In real-world photography, the subject and scene content are often determined by the shooting environment and creative intent, creating a need to improve image composition and visual presentation through fine-grained viewpoint adjustment while preserving the original content. To this end, we introduce **Aesthetic-Guided 6-DoF Image Editing**. The task presents two key challenges: composition optimization typically requires achieving substantial improvement through small viewpoint changes within a limited camera-motion range, while real viewpoint adjustment involves both 3D translation and 3D rotation and thus requires full 6-DoF modeling. However, there is no dataset that couples fine-grained 6-DoF camera changes with substantial composition improvement. Aesthetic priors are particularly valuable for this task, as they indicate which compositional factors most affect visual quality and how they should be adjusted. Reliable supervision should therefore use aesthetic priors to guide the corresponding DoFs and motion directions, while covering the full 6-DoF space of translations, rotations, and their combinations. Based on this insight, we construct **Aes6D**, a dataset and benchmark for this task. Aes6D covers nine composition patterns, provides complete 6-DoF supervision with both *Single* and *Multi* samples, and uses aesthetic guidance to make small viewpoint changes consistently produce substantial and compositionally meaningful differences. Building on Aes6D, we further conduct an initial exploration of structured 6-DoF learning with **DoF-GRPO**, which introduces per-DoF geometric rewards, field-wise relative advantages, and token-level credit assignment to better align GRPO with the geometric structure of 6-DoF camera parameters. Experiments show that models trained on Aes6D achieve substantial composition improvement while preserving scene content, outperforming existing composition and image editing methods and validating Aes6D for this task.

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.

1 INTRODUCTION

Image editing has made substantial progress in modifying subjects, scene content, and visual appearance (Brack et al., 2024; Couairon et al., 2022; Meng et al., 2021; Mokady et al., 2023; Zhang et al., 2025; Avrahami et al., 2025; Kulikov et al., 2025; Wang et al., 2024). However, in real-world photography, the subject and scene content are often determined by the shooting environment and creative intent. In such scenarios, a key need is to **improve composition and visual presentation through viewpoint adjustment without substantially changing the subject, scene structure, or semantic content**. This capability is valuable for capture-time assistance, viewpoint refinement, and camera-controllable editing. Yet existing methods based on cropping, 2D adjustment, or generative modification (Fang et al., 2014; Jia et al., 2022; Li et al., 2020; Su et al., 2024; Tu et al., 2020; Wang & Shen, 2017; Wei et al., 2018; You et al., 2026; Du et al., 2026) often struggle to achieve substantial composition improvement and strong scene consistency.

To address this problem, we introduce a new task: *Aesthetic-Guided 6-DoF Image Editing*. Given an image, the model predicts a fine-grained viewpoint adjustment in the full 6-DoF camera-motion space based on its composition issues, aiming to improve visual presentation. This task presents two key challenges. **1) Fine-Grained Viewpoint Control.** Composition adjustment in real-world photography typically occurs within a limited camera-motion range, aiming to achieve substantial composition improvement through small camera motions (Langford, 2000; Freeman, 2007). Insufficient motion may produce little improvement, while inappropriate adjustment can degrade the original composition, making the required fine-grained control difficult for conventional image editing. **2) Full 6-DoF Modeling.** Real camera motion consists of 3D translation and 3D rotation, and different composition issues may depend on different motion dimensions or coordinated changes across multiple degrees of freedom. Modeling only a subset of these dimensions is therefore insufficient to fully characterize real viewpoint adjustment (Li et al., 2025a; Liao et al., 2026; Bernal-Berdun et al., 2025).

Learning such fine-grained 6-DoF editing requires reliable supervision, yet there is no dataset that couples fine-grained 6-DoF camera changes with substantial composition improvement. Existing data sources remain insufficient. 360° panoramas (Li et al., 2025a; Liao et al., 2026) typically provide only limited camera variations; video data (Ling et al., 2024; Zhou et al., 2018; Knapitsch et al., 2017) do not guarantee reliable high-quality target views; and although 3D-based view synthesis (Jiang et al., 2026; Bai et al., 2026; Liu et al., 2023b; Seo et al., 2024; Shi et al., 2023; Höllein et al., 2024) enables camera-pose changes, unguided motions within a limited range often fail to consistently produce compositionally meaningful differences (Xiao et al., 2026). These limitations suggest two requirements for supervision. First, small random motions often lack specificity to composition issues, whereas aesthetic priors can identify the compositional factors that require adjustment and determine the DoFs and motion directions, enabling limited viewpoint changes to consistently produce substantial composition differences. Second, the supervision should cover the full 6-DoF camera-motion space, including translations, rotations, and coordinated changes across multiple DoFs, thereby covering the full range of viewpoint adjustments required by the task.

Based on this observation, we construct **Aes6D**, a dataset and benchmark for Aesthetic-Guided 6-DoF Image Editing. Starting from high-quality Good Views, we combine photographic composition knowledge with aesthetic priors (Lee et al., 2018; Zhang et al., 2021; Zhao et al., 2025) to cover nine representative composition patterns, spanning most composition cases that can be improved through viewpoint adjustment. Aes6D provides complete 6-DoF supervision and includes both *Single* and *Multi* samples to cover single- and multi-DoF adjustments, balancing clear supervision with motion diversity. Through this aesthetic-guided construction, small viewpoint changes consistently produce substantial and compositionally meaningful differences while preserving scene content.

Building on the complete 6-DoF supervision provided by Aes6D, we further conduct an initial exploration of structured 6-DoF learning with **DoF-GRPO**. Standard GRPO does not account for the geometry of camera parameters, while its shared sequence-level rewards and advantages make it difficult to distinguish prediction quality across different DoFs. DoF-GRPO therefore introduces per-DoF geometric rewards, field-wise relative advantages, and token-level credit assignment as a DoF-aware extension to standard GRPO.

Experiments demonstrate the effectiveness of Aes6D for Aesthetic-Guided 6-DoF Image Editing. Models trained on Aes6D learn fine-grained full 6-DoF viewpoint adjustments that substantially improve composition while preserving the subject and scene content, supporting translations, rotations,

and their combinations. Comparisons with existing composition and image editing methods further demonstrate the utility of Aes6D for learning and evaluating this new task.

In summary, the main contributions of this work are as follows:

- We introduce **Aesthetic-Guided 6-DoF Image Editing**, a new task for real-world photographic composition optimization. It performs fine-grained viewpoint adjustment in the full 6-DoF camera-motion space, aiming to achieve substantial composition improvement through small camera changes while maintaining scene consistency.
- We construct **Aes6D**, a dataset and benchmark for this task. Aes6D combines composition knowledge with aesthetic priors, covers nine composition patterns, and provides complete 6-DoF supervision with *Single* and *Multi* samples, enabling small viewpoint changes to consistently produce substantial and compositionally meaningful differences.
- Building on Aes6D, we conduct an initial exploration of structured 6-DoF learning with **DoF-GRPO**. It introduces per-DoF geometric rewards, field-wise relative advantages, and token-level credit assignment to better adapt GRPO to 6-DoF camera parameters.

2 Related Work

Image Editing. Recent diffusion models (Zhang et al., 2025; Avrahami et al., 2025; Kulikov et al., 2025; Wang et al., 2024; Yoon et al., 2025; Song et al., 2025; Esser et al., 2024; Labs, 2024; Zhou et al., 2026) and unified multimodal models (Liu et al., 2025a; Wang et al., 2025; Wu et al., 2025a; Chen et al., 2025; Cui et al., 2025; Deng et al., 2025; Li et al., 2025b; Lin et al., 2025; Wu et al., 2025b; Xiao et al., 2025) have substantially improved image editing for content, style, and structural modifications. Existing methods are typically driven by textual instructions and primarily focus on instruction following and semantic consistency. However, they mostly operate directly in 2D image space without explicitly modeling viewpoint adjustment in the camera-motion space. In contrast, we formulate aesthetic-guided composition optimization as an image editing task through fine-grained viewpoint adjustment in the full 6-DoF camera-motion space.

Composition Guidance. Existing composition research mainly includes composition understanding (Lee et al., 2018; Zhang et al., 2021; Zhao et al., 2025) and image cropping (Fang et al., 2014; Jia et al., 2022; Li et al., 2020; Su et al., 2024; Tu et al., 2020; Wang & Shen, 2017; Wei et al., 2018). Traditional methods improve composition through aesthetic scoring, candidate-region selection, or cropping-box regression, while recent works (You et al., 2026; Du et al., 2026; Li et al., 2025a; Liao et al., 2026) further explore viewpoint-aware composition guidance and structural reconstruction. However, these methods either rely on 2D image-space transformations, provide coarse viewpoint guidance, or model only a subset of camera-motion dimensions, leaving fine-grained composition optimization in the full 6-DoF space largely unexplored. In contrast, we focus on achieving substantial composition improvement through small camera motions and construct **Aes6D** to provide dedicated supervision for such aesthetic-guided 6-DoF editing.

Aesthetic MLLMs. Recent aesthetic MLLMs (Yang et al., 2026; Cao et al., 2025; 2026; Liu et al., 2025b; Wu et al., 2023; Qi et al., 2025; Huang et al., 2024; Zhou et al., 2024) have made substantial progress in aesthetic scoring and in the perception and reasoning of aesthetics. However, these methods are primarily designed to evaluate and understand aesthetics rather than translate aesthetic knowledge into executable camera adjustments. In contrast, we leverage aesthetic priors to identify composition attributes and determine which camera degrees of freedom should be adjusted and in which directions, connecting aesthetic understanding with 6-DoF image editing.

3 Aes6D: A New Dataset and Benchmark

Fine-grained viewpoint adjustment in Aesthetic-Guided 6-DoF Image Editing occurs within a limited motion range, where small viewpoint changes produce noticeable aesthetic differences. The challenge is to construct view pairs where small 6-DoF changes consistently produce substantial, compositionally meaningful differences while preserving scene content and motion supervision.

To address this challenge, we introduce **Aes6D**, a dataset and benchmark for fine-grained 6-DoF composition optimization. Its construction is guided by a principle: camera motions should induce

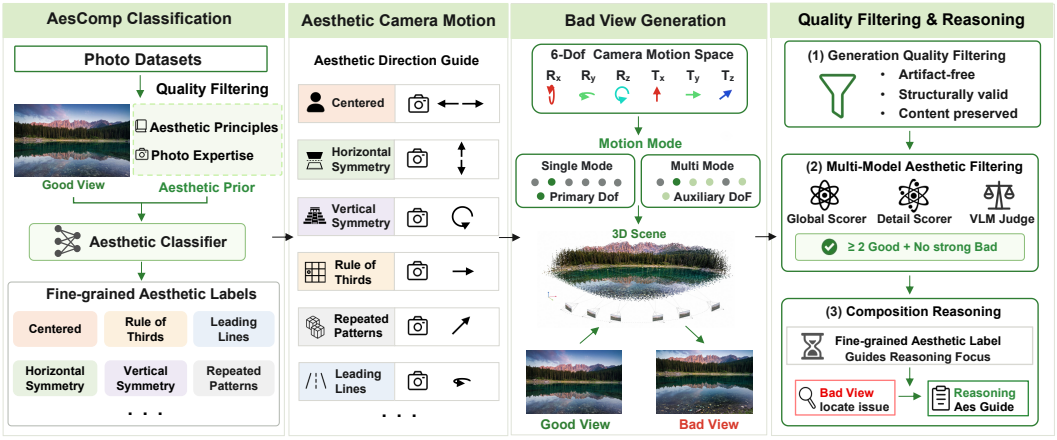


Figure 2: Overview of the Aes6D construction pipeline. Starting from high-quality Good Views, we identify composition categories and attributes, translate aesthetic principles into directional 6-DoF camera motions to generate Bad Views, and apply quality filtering and composition reasoning.

clear and compositionally meaningful changes while remaining small enough to preserve scene content. We start from high-quality expert-taken photos as *Good Views* and translate their composition principles into directional 6-DoF camera motions to generate *Bad Views*. To capture single-DoF and coordinated camera adjustments, we construct two sample types, *Single* and *Multi*, corresponding to single-DoF and multi-DoF motions, respectively. Since the applied motions are known, their inverse motions provide 6-DoF supervision for recovering the Good View from the Bad View. Meanwhile, based on the disrupted composition principle, we describe the composition issue in the Bad View and its adjustment direction, providing composition-reasoning supervision. Each sample contains a Good–Bad view pair, composition reasoning, and complete 6-DoF camera-motion supervision. Aes6D contains 26,834 samples in total, including 25,834 training samples and 1,000 benchmark samples. Representative examples are provided in Appendix A, and the overall construction pipeline is shown in Figure 2. Implementation details are provided in Appendix C.

3.1 Good View Collection and Aesthetic Composition Classification

We select high-quality photographs from TAD66K (He et al., 2022), AVA (Murray et al., 2012), and Unsplash Lite (Unsplash, 2023), and further filter them using aesthetic assessment models to ensure strong photographic quality and reliable composition.

To guide subsequent camera-motion design, we first annotate each Good View with fine-grained **composition attributes**. Based on photography knowledge and visual organization principles from prior aesthetic studies, we summarize nine representative composition categories: *centered composition, rule of thirds, horizontal symmetry, vertical symmetry, horizontal structure, vertical structure, leading lines, repeated patterns, and artistic composition*. These categories capture distinct compositional patterns that can be translated into directional camera adjustments.

For directional composition categories, we further annotate **spatial attributes**. For example, rule-of-thirds records the subject region, horizontal structure the dominant structure location, and leading lines the dominant line location and direction.

3.2 Aesthetic-Guided Camera Motion and Bad View Generation

Within a limited camera-motion range, random camera motions often fail to produce clear and consistent composition differences. This is because they ignore the aesthetic composition principles underlying the Good View, which provide explicit cues for viewpoint adjustment.

We therefore use the composition categories and spatial attributes obtained in the previous stage to guide camera motion in a unified full 6-DoF space. Different composition principles correspond to different dominant motion directions. For example, compositions governed by subject placement

are mainly adjusted through horizontal translation or yaw; horizontal structures through vertical translation or pitch; and vertical structures through roll. We thus translate photographic composition principles into directional constraints in the 6-DoF camera-motion space, enabling small viewpoint changes to produce clearer and more compositionally meaningful differences.

To cover camera adjustments of varying complexity, we introduce two motion modes, *Single* and *Multi*. A Single sample contains one primary DoF, while a Multi sample combines primary and auxiliary motions. We then apply these motions through a view-synthesis pipeline (Wu et al., 2025a; Bochkovskiy et al., 2025) and render the corresponding Bad Views from the resulting camera poses.

3.3 Quality Filtering and Composition Reasoning

Automatically generated Bad Views may contain completion artifacts, structural abnormalities, or insufficient composition differences. We employ a two-stage quality-control process. We first perform generation-quality filtering, using a vision-language model (Qwen Team, 2026b) to remove samples with subject loss, structural corruption, completion artifacts, or unnatural content. We then perform multi-model aesthetic filtering, combining a general aesthetic score (Cao et al., 2026), fine-grained aesthetic assessment (Yang et al., 2026), and VLM-based Good–Bad pairwise comparison (Qwen Team, 2026b) to verify compositional differences between Good and Bad Views. A sample is retained if at least two evaluators prefer the Good View and none strongly favors the Bad View.

In addition to 6-DoF camera parameters, we generate composition reasoning for each sample (Qwen Team, 2026b). Since each Bad View is obtained by directionally disrupting an identified composition principle in its Good View, we use this relationship to describe the composition issue visible in the Bad View and its corresponding direction of improvement, keeping the reasoning aligned with both the underlying composition principle and the applied camera motion.

3.4 Benchmark Construction

We select 1,000 samples from the final dataset to form the Aes6D benchmark and use the remaining samples for training. To ensure diverse and balanced coverage, we perform stratified sampling over Single/Multi-motion modes, composition categories, and scene categories, covering diverse composition issues, spatial adjustment patterns, and real-world photographic scenes.

We further conduct human quality review on the benchmark to remove samples with generation artifacts, ambiguous Good–Bad composition relationships, or other issues that may affect evaluation reliability, reducing residual noise from the automatic construction pipeline.

4 Method

Our goal is to predict 6-DoF motion from composition issues in the input view for composition improvement. As an initial exploration, we adopt a two-stage training framework. We first perform supervised fine-tuning (Liu et al., 2023a) to learn basic structured 6-DoF camera control, and then apply **DoF-GRPO** with per-DoF geometric rewards, field-wise relative advantages, and token-level credit assignment to refine 6-DoF predictions. The overall framework is illustrated in Figure 3.

4.1 Basic 6-DoF Control

Given a Bad View I_b and a task prompt q, the model first generates composition reasoning r describing the composition issue and its direction of improvement, followed by a structured 6-DoF camera motion m. The complete output is formulated as:

$$y = [r, m]. \tag{1}$$

Here, r denotes the composition reasoning, while m specifies camera changes along three rotational and three translational degrees of freedom.

We train the model on Aes6D using the standard autoregressive objective:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{t=1}^{|y|} \log \pi_\theta \left(y_t \mid I_b, q, y_{<t} \right). \tag{2}$$

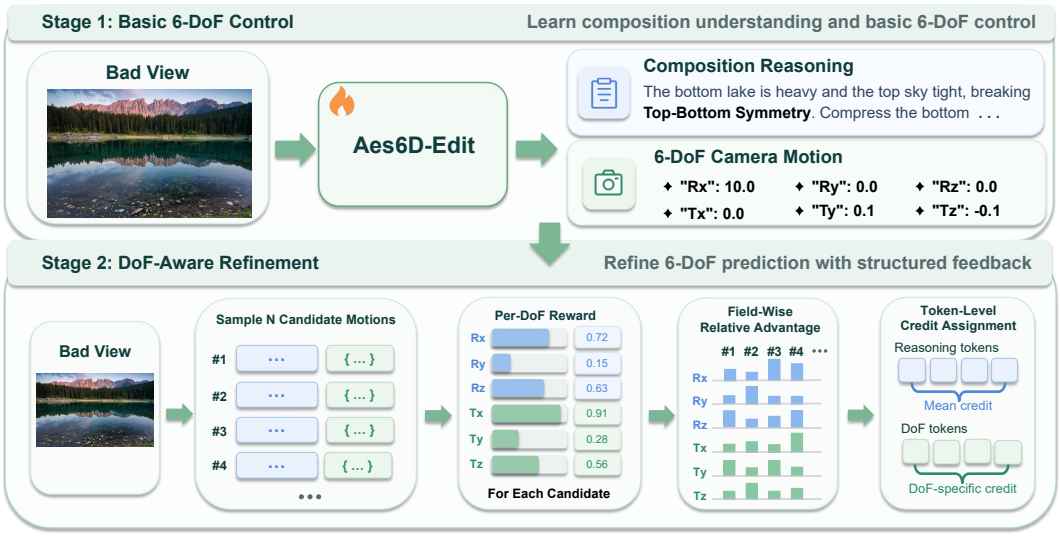


Figure 3: Overview of our two-stage training framework. SFT first learns composition reasoning and structured 6-DoF camera control from supervised examples. DoF-GRPO then computes per-DoF rewards for each candidate, derives field-wise relative advantages across the sampled group, and assigns DoF-specific or mean credit to the corresponding output tokens for fine-grained optimization.

SFT establishes the mapping from composition understanding to structured 6-DoF camera control, enabling the model to jointly generate composition reasoning and camera parameters.

4.2 DoF-Aware Refinement

Although SFT establishes basic 6-DoF control, its token-likelihood objective does not account for the geometric relationships of camera parameters (Kendall & Cipolla, 2017; Diaz & Marathe, 2019; Kim et al., 2024). For example, 0° and 359° are numerically far apart but differ by only 1° in orientation. Standard GRPO (Shao et al., 2024) is also ill-suited to structured 6-DoF prediction, as shared sequence-level rewards and advantages cannot distinguish prediction quality across DoFs or accurately assign feedback to the corresponding numerical tokens.

To address these issues, we propose **DoF-GRPO** (*Degree-of-Freedom-aware GRPO*), which performs geometry-aware per-DoF rewards, field-wise relative advantages, and token-level credit assignment for fine-grained optimization of structured 6-DoF predictions.

Geometry-Aware DoF Reward. For the k-th degree of freedom of candidate i, we independently measure the error between the prediction $\hat{m}_{i,k}$ and the reference m_k^*. Since rotational parameters are periodic, we use the shortest circular distance for rotation and absolute error for translation:

$$e_k(\hat{m}_{i,k}, m_k^*) = \begin{cases} d_{\text{circ}}(\hat{m}_{i,k}, m_k^*), & k \in \mathcal{K}_{\text{rot}}, \\ |\hat{m}_{i,k} - m_k^*|, & \text{otherwise}, \end{cases} \tag{3}$$

where

$$d_{\text{circ}}(a, b) = \min\left(|a - b| \bmod 360,\ 360 - (|a - b| \bmod 360)\right). \tag{4}$$

We then define a per-DoF reward as

$$r_{i,k} = \exp\left(-\frac{e_k(\hat{m}_{i,k}, m_k^*)}{s_k}\right), \tag{5}$$

where $s_k > 0$ is a DoF-specific scale determined by the motion range of the k-th dimension.

Field-Wise Relative Advantage. For the k-th DoF of candidate i, we normalize its reward $r_{i,k}$ based on the reward distribution of the same DoF across all candidates in the sampled group:

$$A_{i,k} = \frac{r_{i,k} - \mu_k}{\sigma_k + \epsilon}, \tag{6}$$

Table 1: Quantitative comparison on the **Aes6D** benchmark. Our method achieves the best overall performance. Best results are highlighted in **bold**, and second-best results are underlined.

Model	Win Rate vs. Bad View ↑		Win Rate vs. Good View ↑		Scene Consistency Rate ↑	
	GPT-5	Human	GPT-5	Human	GPT-5	Human
BAGEL	12.0%	15.0%	3.0%	7.0%	45.0%	57.0%
FLUX.1 Kontext	35.0%	32.0%	8.0%	12.0%	93.0%	89.0%
Qwen-Image-Edit-2511	29.0%	23.0%	6.0%	11.0%	48.0%	47.0%
PhotoFramer	49.0%	57.0%	34.0%	37.0%	80.0%	66.0%
AesFormer	39.0%	21.0%	20.0%	13.0%	14.0%	9.0%
Ours	**81.0%**	**83.0%**	**54.0%**	**60.0%**	**94.0%**	**93.0%**

where μ_k and σ_k denote the mean and standard deviation of the k-th DoF rewards across all candidates in the group, respectively, and ϵ is a small constant for numerical stability.

Each candidate therefore obtains six independent DoF-level advantages, reflecting the relative prediction quality of each motion dimension within the group.

Token-Level Credit Assignment. To further align the DoF-level feedback with the structured model output, we locate the token span corresponding to each numerical 6-DoF field and assign the corresponding $A_{i,k}$ to those tokens. Other tokens, including composition reasoning and JSON structure, receive the mean advantage across all six degrees of freedom:

$$A_{i,t}^{\text{tok}} = \begin{cases} A_{i,k}, & t \in \text{span}(k), \\ \frac{1}{6}\sum_{k=1}^{6} A_{i,k}, & \text{otherwise.} \end{cases} \tag{7}$$

where $\text{span}(k)$ denotes the token positions of the k-th DoF value.

In this way, the sequence-level feedback in standard GRPO is converted into token-level credit assignment aligned with the structured 6-DoF output.

We use a PPO-style clipped objective (Schulman et al., 2017) with token-level advantages:

$$\mathcal{L}_{\text{DoF-GRPO}} = -\mathbb{E}_{i,t}\left[\min\left(\rho_{i,t}A_{i,t}^{\text{tok}}, \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)A_{i,t}^{\text{tok}}\right)\right] + \beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right), \tag{8}$$

where

$$\rho_{i,t} = \frac{\pi_\theta(y_{i,t} \mid \cdot)}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid \cdot)} \tag{9}$$

is the probability ratio between the current and old policies, and the KL term (Kullback & Leibler, 1951) constrains the policy from drifting excessively from the SFT model.

5 Experiments

5.1 Implementation Details

Training. We adopt Qwen3.5-9B (Qwen Team, 2026a) as the base model and perform LoRA-based (Hu et al., 2021) SFT on Aes6D for 5 epochs with a learning rate of 1×10^{-4} and an effective batch size of 128. Given a Bad View, the model predicts composition reasoning that identifies the composition issue and improvement direction, together with the corresponding 6-DoF camera motion. The predicted camera motion is then applied through the view-synthesis pipeline to render the final edited view. We further fine-tune the SFT model with DoF-GRPO using 8 rollouts per input and a learning rate of 1×10^{-6}. All experiments are conducted on 8 NVIDIA A100 GPUs.

Evaluation. We sample 10% of the Aes6D benchmark and evaluate methods with GPT-5 (Singh et al., 2025) and human annotators under the same protocol. We report **Win Rate vs. Bad View**, **Win Rate vs. Good View**, and **Scene Consistency Rate**, measuring preference over the Bad View, preference over the Good View, and scene consistency with the Bad View, respectively. We evaluate on 50 web-collected OOD images outside the Aes6D construction pipeline, reporting Win Rate vs.

Table 2: Ablation studies on Aes6D, analyzing the contributions of training supervision and DoF-GRPO optimization. Best results are highlighted in **bold**.

Setting	Win Rate vs. Bad View ↑		Win Rate vs. Good View ↑		Scene Consistency Rate ↑	
	GPT-5	Human	GPT-5	Human	GPT-5	Human
Single only	57.0%	59.0%	46.0%	50.0%	91.0%	92.0%
Multi only	64.0%	64.0%	50.0%	53.0%	85.0%	87.0%
Single + Multi	**81.0%**	**83.0%**	**54.0%**	**60.0%**	**94.0%**	**93.0%**
SFT only	74.0%	72.0%	47.0%	51.0%	**94.0%**	91.0%
SFT + Standard GRPO	73.0%	76.0%	51.0%	55.0%	93.0%	92.0%
SFT + DoF-GRPO	**81.0%**	**83.0%**	**54.0%**	**60.0%**	**94.0%**	**93.0%**

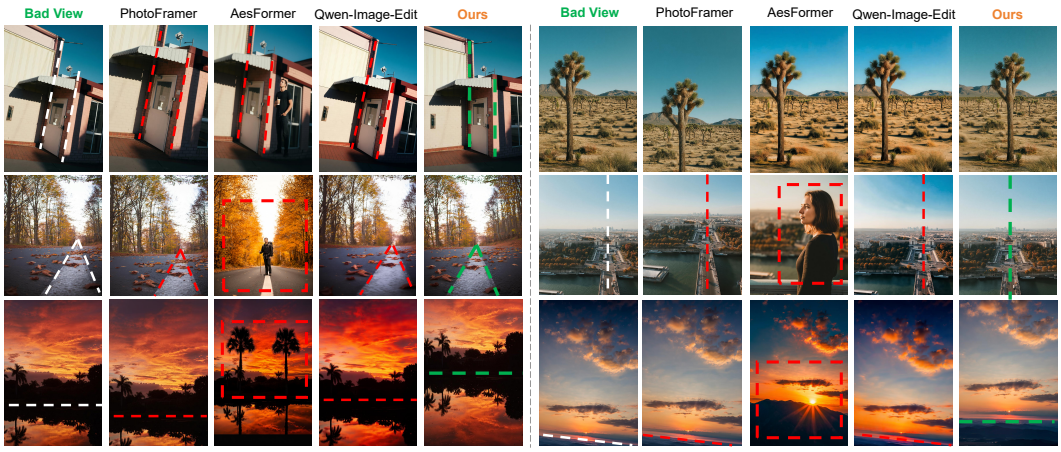


Figure 4: Qualitative comparison on Aes6D. Our method improves composition through diverse 6-DoF viewpoint adjustments while retaining the original scene content.

Bad View and Scene Consistency Rate without Good Views. We explore iterative editing under consecutive viewpoint adjustments in Appendix D. We do not report PSNR/SSIM, as pixel-level similarity is not well-suited to evaluating viewpoint changes, and we avoid absolute aesthetic scores because existing aesthetic models are less reliable for fine-grained quality assessment.

5.2 Main Results

Baselines. We compare our method with BAGEL (Deng et al., 2025), FLUX.1 Kontext (Labs et al., 2025), Qwen-Image-Edit-2511 (Wu et al., 2025a), PhotoFramer (You et al., 2026), and AesFormer (Du et al., 2026). All methods take the same Bad View as input and generate composition-enhanced results following their official inference settings.

Quantitative Results. As shown in Table 1, existing methods often struggle to balance composition improvement and scene consistency, whereas our method achieves the best overall performance under both GPT-based and human evaluation. These results show that our method can effectively improve composition through 6-DoF viewpoint adjustment while maintaining scene consistency. Additional OOD quantitative results are provided in Appendix B.

Qualitative Results. Figure 4 shows qualitative comparisons across methods. Existing approaches either provide limited composition improvement or alter the subject or scene content, while our method achieves substantial improvement through fine-grained viewpoint adjustment with better scene consistency. As shown in Figure 5, similar results are observed on web-collected OOD images, indicating a certain degree of generalization beyond the Aes6D distribution.

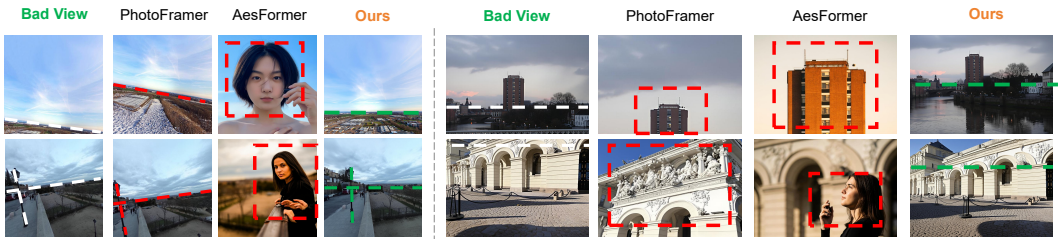


Figure 5: Qualitative results on web-collected OOD images. Our method improves composition while preserving scene content, showing a certain degree of generalization.



Figure 6: Comparison of Random and Aesthetic-Guided camera motions. Aesthetic guidance produces more targeted and compositionally meaningful viewpoint changes.



Figure 7: Qualitative ablation of DoF-GRPO. DoF-GRPO enables more accurate and effective 6-DoF viewpoint adjustments than SFT.

5.3 Ablation Studies

Aesthetic-Guided Camera Motion. As shown in Figure 6, within a limited motion range, random camera motions often produce weak or compositionally irrelevant changes, whereas aesthetic guidance yields more targeted and compositionally meaningful viewpoint changes, validating the benefit of incorporating composition principles into camera-motion design.

Single- and Multi-DoF Supervision. We compare **Single only**, **Multi only**, and **Single + Multi** modes. As shown in Table 2, their combination achieves the best overall performance. Single-DoF provides clearer supervision for individual motions, while Multi-DoF covers more complex viewpoint adjustments, exposing the model to spatial adjustment patterns from isolated to coordinated motions.

DoF-GRPO Optimization. As shown in Table 2, standard GRPO yields modest but inconsistent gains over SFT, whereas DoF-GRPO achieves higher win rates in both GPT-5 and human evaluation while maintaining strong Scene Consistency. This demonstrates the benefit of DoF-aware optimization for structured 6-DoF control. Figure 7 provides qualitative comparisons with SFT.

6 Conclusion

In this work, we introduce **Aesthetic-Guided 6-DoF Image Editing**, aiming for substantial composition improvement through small, fine-grained viewpoint changes in the full 6-DoF space. To support this task, we construct **Aes6D**, which uses aesthetic priors to guide camera motion, covers nine composition patterns, and provides complete 6-DoF supervision with *Single* and *Multi* samples. Building on Aes6D, we further explore structured 6-DoF learning with **DoF-GRPO**.

AI USE STATEMENT

In this work, we used generative AI tools to assist with paper writing and language polishing, as well as for research ideation and auxiliary discussion during research execution, including method design, experimental setup, and result presentation. Generative models were also used as components of the Aes6D construction pipeline for view completion and refinement, as well as composition-reasoning generation. All AI-assisted outputs were validated through the data filtering and quality-control procedures described in the paper and reviewed by the authors.

We did not use generative AI to fabricate or alter reported experimental results or to generate mathematical proofs. The core research questions, method design, experimental implementation, result analysis, and final conclusions were determined by the authors. We reviewed all AI-assisted content and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the Aes6D construction pipeline, the two-stage training framework, and the DoF-GRPO optimization procedure in the main paper. The experimental section reports the main training configurations and evaluation protocol, while representative dataset examples are provided in the appendix and supplementary materials. We also plan to publicly release Aes6D to facilitate further research and reproducibility.

REFERENCES

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 7877–7888, June 2025.

Yunpeng Bai, Haoxiang Li, and Qixing Huang. Positional encoding field. In *International Conference on Learning Representations*, volume 2026, pp. 84996–85008, 2026.

Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Geoffroy, Xin Sun, and Diego Gutierrez. Precisecam: Precise camera control for text-to-image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2724–2733. IEEE, 2025.

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, volume 2025, pp. 75602–75637, 2025.

Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8861–8870, 2024.

Shuo Cao, Jiayang Li, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin, Yu Zhou, et al. Unipercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. *arXiv preprint arXiv:2512.21675*, 2025.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15313–15322, 2026.

Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, et al. Blip3o-next: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners, 2025. URL https://arxiv.org/abs/2510.26583.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4733–4742. IEEE, 2019.

Tianxiang Du, Hulingxiao He, and Yuxin Peng. Aesformer: Transform everyday photos into beautiful memories. *arXiv preprint arXiv:2605.22126*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Xiangyu Fan, Zesong Qiu, Fanzhou Wang, Peng Gao, and Lei Yang. Qwen-image-lightning, 2025. URL https://github.com/ModelTC/LightX2V-Qwen-Image-Lightning.

Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1105–1108, 2014.

Michael Freeman. *The photographer's eye: Composition and design for better digital photos*. CRC Press, 2007.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *IJCAI*, volume 3, pp. 6, 2022.

Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. Viewdiff: 3d-consistent image generation with text-to-image models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5043–5052. IEEE, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5911–5920, 2024.

Gengyun Jia, Huaibo Huang, Chaoyou Fu, and Ran He. Rethinking image cropping: Exploring diverse compositions from global views. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2436–2445. IEEE, 2022.

Hong Jiang, Wensong Song, Zongxin Yang, Ruijie Quan, and Yi Yang. Unigeo: Unifying geometric guidance for camera-controllable image editing via video models. *arXiv preprint arXiv:2604.17565*, 2026.

Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6555–6564. IEEE, 2017.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Michael Langford. *Basic photography*. Taylor & Francis, 2000.

Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Photographic composition classification and dominant geometric element detection for outdoor scenes. *Journal of Visual Communication and Image Representation*, 55:91–105, 2018.

Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4212–4221. IEEE, 2020.

Jiawan Li, Fei Zhou, Zhipeng Zhong, Jiongzhi Lin, and Guoping Qiu. Towards smart point-and-shoot photography. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28242–28251. IEEE, 2025a.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025b.

Kang Liao, Size Wu, Zhonghua Wu, Linyi Jin, Chao Wang, Yikai Wang, Fei Wang, Wei Li, and Chen Change Loy. Thinking with camera: A unified multimodal model for camera-centric understanding and generation. In *International Conference on Learning Representations*, volume 2026, pp. 144339–144375, 2026.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023b.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025a.

Yuti Liu, Shice Liu, Junyuan Gao, Peng-tao Jiang, Hao Zhang, Jinwei Chen, and Bo Li. Advancing comprehensive aesthetic insight with multi-scale text-guided self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5748–5756, 2025b.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Lars Mescheder, Wei Dong, Shiwei Li, Xuyang Bai, Marcel Santos, Peiyun Hu, Bruno Lecouat, Mingmin Zhen, Amaël Delaunoy, Tian Fang, et al. Sharp monocular view synthesis in less than a second. In *International Conference on Learning Representations*, volume 2026, pp. 24192–24230, 2026.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2408–2415. IEEE, 2012.

Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. The photographer's eye: Teaching multimodal large language models to see and critique like photographers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24807–24816. IEEE, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL `https://qwen.ai/blog?id=qwen3.5`.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *Advances in Neural Information Processing Systems*, 37:80220–80243, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Wensong Song, Hong Jiang, Zongxing Yang, Ruijie Quan, and Yi Yang. Insert anything: Image insertion via in-context editing in dit. *arXiv preprint arXiv:2504.15009*, 2025.

Yukun Su, Yiwen Cao, Jingliang Deng, Fengyun Rao, and Qingyao Wu. Spatial-semantic collaborative cropping for user generated content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4988–4997, 2024.

Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12104–12111, 2020.

Unsplash. Unsplash dataset. `https://unsplash.com/data`, 2023. Accessed: 2023-12-15.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Jianshan Zhao, Yang Li, and Qing-Guo Chen. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

Wenguan Wang and Jianbing Shen. Deep cropping via attention box prediction and aesthetics assessment. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2205–2213. IEEE, 2017.

Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5437–5446. IEEE, 2018.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Yicheng Xiao, Wenhu Zhang, Lin Song, Yukang Chen, Wenbo Li, Nan Jiang, Tianhe Ren, Haokun Lin, Wei Huang, Haoyang Huang, et al. Spatialedit: Benchmarking fine-grained image spatial editing. *arXiv preprint arXiv:2604.04911*, 2026.

Zhichao Yang, Jianjie Wang, Zhixianhe Zhang, Pangu Xie, Xiangfei Sheng, Pengfei Chen, and Leida Li. Fine-grained image aesthetic assessment: Learning discriminative scores from relative ranks. *arXiv preprint arXiv:2603.03907*, 2026.

Sung-Hoon Yoon, Minghan Li, Gaspard Beaudouin, Congcong Wen, Muhammad Rafay Azhar, and Mengyu Wang. Splitflow: Flow decomposition for inversion-free text-to-image editing. *arXiv preprint arXiv:2510.25970*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10197–10207, 2026.

Bo Zhang, Li Niu, and Liqing Zhang. Image composition assessment with saliency-augmented multi-pattern pooling. *arXiv preprint arXiv:2104.03133*, 2021.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large-scale diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.20690.

Zhaoran Zhao, Peng Lu, Anran Zhang, Peipei Li, Xia Li, Xuannan Liu, Yang Hu, Shiyi Chen, Liwei Wang, and Wenhao Guo. Can machines understand composition? dataset and benchmark for photographic image composition embedding and understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14411–14421. IEEE, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Zhaokun Zhou, Qiulin Wang, Bin Lin, Yiwei Su, Rui Chen, Xin Tao, Amin Zheng, Li Yuan, Pengfei Wan, and Di Zhang. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*, 2024.

Zhenglin Zhou, Fan Ma, Chengzhuo Gui, Xiaobo Xia, Hehe Fan, Yi Yang, and Tat-Seng Chua. Anchorflow: Training-free 3d editing via latent anchor-aligned flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14387–14397, 2026.

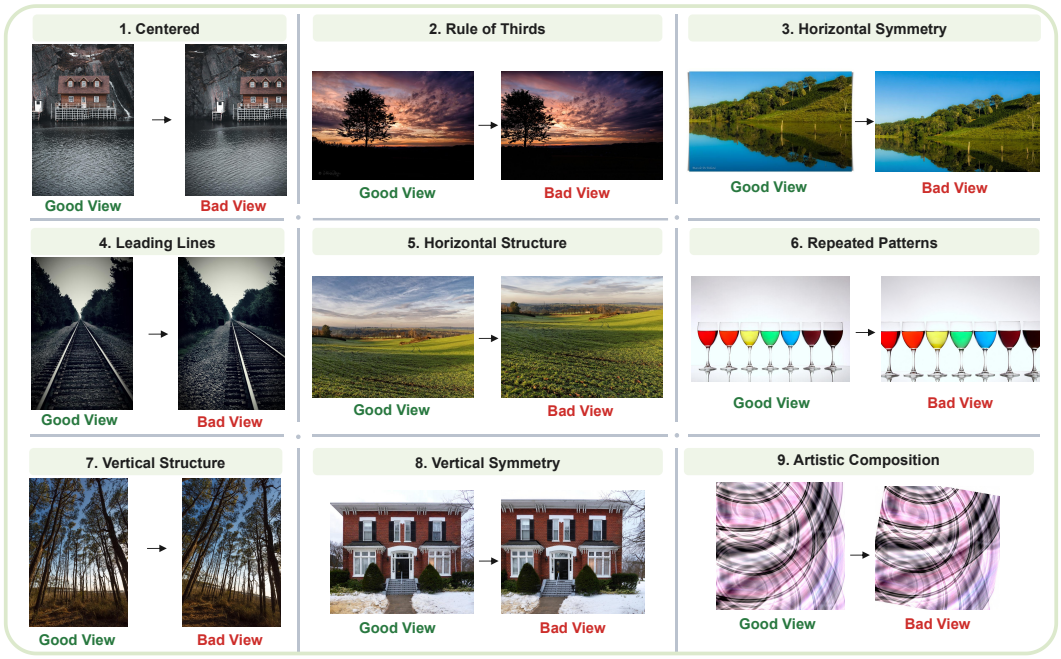


Figure 8: Representative Good–Bad view pairs from Aes6D across nine composition categories, generated through aesthetic-guided 6-DoF camera motion.

APPENDIX

This appendix provides supplementary materials to complement the main manuscript. Section A presents representative examples from Aes6D across different composition categories. Section B provides additional quantitative evaluation on 50 web-collected OOD images outside the Aes6D construction pipeline. Section C provides additional implementation details of the Aes6D construction pipeline. Section D presents iterative refinement under consecutive viewpoint adjustments. Finally, Sections E and F discuss future directions and broader impacts, risks, and limitations, respectively.

A ADDITIONAL AES6D EXAMPLES

Figure 8 presents representative Good–Bad view pairs from Aes6D across the nine composition categories used in our dataset. These examples illustrate how aesthetic-guided camera motions produce clear and compositionally meaningful changes within a limited motion range while largely preserving the original scene content. The examples also cover both single-DoF and coordinated multi-DoF adjustments, reflecting the diversity of spatial supervision provided by Aes6D.

B OOD EVALUATION

To further evaluate generalization beyond the Aes6D construction distribution, we collect an additional set of **50 Internet images** outside the Aes6D construction pipeline. This evaluation examines whether the composition understanding and 6-DoF viewpoint adjustment learned from Aes6D can transfer to unseen image distributions.

We follow the same GPT-5 and human pairwise evaluation protocol as in the main paper and report **Win Rate vs. Bad View** and **Scene Consistency Rate**. As shown in Table 3, our method achieves the best overall performance in both composition improvement and scene consistency, indicating a certain degree of generalization beyond the Aes6D construction distribution.

Table 3: Quantitative comparison on 50 web-collected OOD images outside the Aes6D construction pipeline. Best results are highlighted in **bold**, and second-best results are underlined.

Model	Win Rate vs. Bad View ↑		Scene Consistency Rate ↑	
	GPT-5	Human	GPT-5	Human
BAGEL	10.0%	12.0%	60.0%	66.0%
FLUX.1 Kontext	14.0%	18.0%	90.0%	90.0%
Qwen-Image-Edit-2511	36.0%	34.0%	62.0%	58.0%
PhotoFramer	42.0%	54.0%	70.0%	62.0%
AesFormer	40.0%	30.0%	24.0%	16.0%
Ours	**64.0%**	**56.0%**	**92.0%**	**96.0%**

Table 4: Composition-aware camera-motion design in Aes6D.

Composition	Primary DoF	Auxiliary DoFs
Centered	T_x / R_y	T_y, T_z, R_x, R_z
Rule of Thirds	T_x / R_y	T_y, T_z, R_x, R_z
Horizontal Symmetry	T_y / R_x	T_x, T_z, R_y, R_z
Vertical Symmetry	T_x / R_y	T_y, T_z, R_x, R_z
Horizontal Structure	T_y / R_x	T_x, T_z, R_y, R_z
Vertical Structure	R_z	R_x, R_y, T_x, T_y, T_z
Leading Lines	T_x / R_y	T_y, T_z, R_x, R_z
Repeated Patterns	T_z	R_x, R_y, R_z, T_x, T_y
Artistic Composition	Any DoF	Remaining DoFs

C Implementation Details of Aes6D

This section provides additional implementation details of Aes6D, including camera-motion parameterization, aesthetic-guided motion design, Single/Multi sampling, and the view-synthesis pipeline used to generate Bad Views.

C.1 Camera Motion Parameterization

We represent the corrective camera motion from a generated Bad View toward its reference Good View as

$$\Delta\mathbf{p} = (R_x, R_y, R_z, T_x, T_y, T_z), \tag{10}$$

where R_x, R_y, and R_z denote pitch, yaw, and roll, respectively, and T_x, T_y, and T_z denote horizontal, vertical, and depth translations.

Rotations are measured in degrees, while translations are represented as normalized ratios. In particular, the in-plane translations T_x and T_y are normalized by the image width and height, respectively. For example, $T_x = 0.10$ corresponds to a horizontal displacement equivalent to 10% of the image width, while $T_y = 0.10$ corresponds to 10% of the image height. During data construction, we apply camera motion to a Good View to generate a Bad View and record the inverse motion as the 6-DoF supervision for recovering the Good View.

C.2 Aesthetic-Guided Motion Design

For each Good View, we first identify its composition category and, when applicable, the location of the dominant subject or structure. We then select a composition-dependent primary DoF, with its direction determined by the corresponding composition attributes. Table 4 summarizes the mapping used in Aes6D. For Multi-DoF samples, additional motions are sampled from the corresponding auxiliary DoFs.

Here, *Horizontal Symmetry* denotes top–bottom symmetry about the horizontal image axis, while *Vertical Symmetry* denotes left–right symmetry about the vertical image axis.

For categories with positional attributes, the motion sign is determined by the annotated subject or structure location. For example, for a rule-of-thirds composition with the subject in the left-third region, we select the horizontal direction that moves it away from this preferred region when generating the Bad View. Similarly, for a horizontal structure located in the upper or lower region, the vertical motion direction is chosen to displace it from its original compositional position. For categories without an explicit directional preference, the motion direction is randomly sampled.

C.3 Single- and Multi-DoF Sampling

A Single sample contains exactly one active DoF,

$$\|\Delta \mathbf{p}\|_0 = 1, \tag{11}$$

whose type is determined by the composition-to-motion mapping above. For Artistic Composition, the active DoF is uniformly sampled from all six dimensions.

For Multi samples, we select a composition-dependent primary DoF together with additional auxiliary DoFs to form multi-DoF camera adjustments.

We use discrete motion magnitudes to produce controlled viewpoint changes:

$$|R_{\text{primary}}| \in \{5^\circ, 10^\circ\}, \qquad |T_{\text{primary}}| \in \{0.10, 0.15, 0.20\}. \tag{12}$$

Auxiliary pitch and yaw use a magnitude of 5°, auxiliary roll uses $\{5^\circ, 10^\circ\}$, and auxiliary translations use $\{0.10, 0.15\}$. The sign of the primary motion follows the direction determined in Sec. C.2, while auxiliary motion directions are sampled independently; the values above specify only the motion magnitude.

C.4 View-Synthesis-Based Bad-View Generation

For Aes6D construction, we use a 3D-aware view-synthesis pipeline. Given a Good View, we first resize the image to the closest aspect-ratio-preserving resolution among

$$\{480\times768,\ 512\times720,\ 608\times608,\ 720\times512,\ 768\times480\}. \tag{13}$$

We then use Depth Pro (Bochkovskiy et al., 2025) to estimate metric depth and focal length from the input image and back-project the image into a colored 3D point cloud representing the observed scene geometry. Points around strong depth discontinuities are removed to reduce stretching artifacts during novel-view rendering. The sampled camera motion is then applied to this scene representation, and the target view is rendered using PyTorch3D.

Viewpoint changes naturally expose regions unobserved in the source image. We define the missing-region ratio as

$$r_{\text{mask}} = \frac{|\mathcal{M}_{\text{miss}}|}{HW}, \tag{14}$$

where $\mathcal{M}_{\text{miss}}$ denotes the set of pixels in the rendered target view that are not covered by the back-projected source point cloud, and H and W denote the target-view height and width. We retain only candidate viewpoints satisfying

$$0.05 \le r_{\text{mask}} \le 0.30. \tag{15}$$

A smaller ratio typically produces insufficient viewpoint variation, whereas a larger ratio requires excessive synthesis of unobserved content. When the constraint is not satisfied, we resample the camera motion for at most 20 trials.

We further refine the rendered view using Qwen-Image-Edit-2511 (Wu et al., 2025a) equipped with Gaussian-Splash (Mescheder et al., 2026) and Lightning LoRAs (Fan et al., 2025). The editor receives both the rendered target view and the original Good View, using the latter as a scene-content reference to complete newly exposed regions and correct rendering artifacts. Both LoRA adapters use a weight of 1.0, and we perform 8 inference steps with deterministic sampling. The refined result is then passed to the subsequent generation-quality and aesthetic filtering stages.

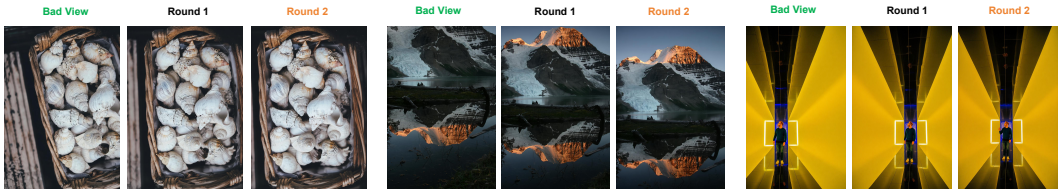


Figure 9: Examples of iterative viewpoint refinement. Larger initial adjustments address major composition issues, followed by smaller refinements for further composition improvement.

D Iterative Spatial Refinement

Although our model is trained for single-step viewpoint adjustment, we further explore its behavior under consecutive refinements. As shown in Figure 9, we feed the result of the first adjustment back into the model for a second prediction. In both examples, the model first applies a relatively larger camera motion to address the main composition issue, followed by a smaller adjustment for further refinement. This suggests that the model can naturally perform progressive viewpoint refinement according to the current composition state.

E Future Directions

While our work focuses on fine-grained composition optimization through local 6-DoF camera adjustment, several directions remain for future exploration.

First, extending the framework beyond photographic composition toward more general spatial editing objectives is a promising direction. Spatially-aware image editing may involve broader spatial relationships among objects, viewpoints, and scene structures. Developing supervision and control mechanisms for these objectives could further expand the scope of spatial image editing.

Second, improving the robustness of fine-grained spatial control across more diverse and complex scenes remains important. Subtle camera adjustments may interact differently with scene depth, object layout, and occlusion patterns, making consistent spatial control and scene preservation challenging. Stronger spatial modeling could improve the reliability of local 6-DoF adjustment across varied real-world scenes.

Third, explicit multi-step spatial control is another promising direction. Future work could investigate iterative policies that repeatedly assess the current view, predict a refinement, and determine when further adjustment is unnecessary. Such a closed-loop formulation may enable more adaptive and precise spatial editing.

F Broader Impacts

F.1 Positive Societal Impacts

Assisting Photographic Composition. Our method provides interpretable camera-motion guidance for improving image composition and may assist photographers in refining viewpoints during image capture. By expressing spatial adjustments through explicit 6-DoF camera motion, the framework offers a direct connection between visual composition and physical camera control.

Supporting Spatially Controllable Image Editing. The proposed framework enables composition improvement primarily through viewpoint adjustment rather than unconstrained modification of scene content. This may provide artists and content creators with a more controllable way to refine spatial organization while maintaining the original scene.

Facilitating Research on Spatial Image Editing. Aes6D provides supervision that connects composition changes with explicit camera motion. We hope that the dataset and benchmark can support further research on spatial understanding, camera control, and spatially-aware image editing.

F.2 Potential Risks and Limitations

Fabricated Viewpoints. Like other generative image editing methods, our approach can synthesize viewpoints that were not physically captured. Such results could potentially be presented as genuine photographs if their synthetic nature is not disclosed.

Hallucination of Unobserved Content. Viewpoint adjustment may reveal regions not visible in the original image, requiring the underlying generative model to synthesize new content. These regions may not correspond to the actual scene and should therefore not be interpreted as faithful reconstruction of unobserved environments.

Bias in Aesthetic Guidance. The composition principles and aesthetic priors used in Aes6D reflect particular photographic conventions and may not capture the full diversity of aesthetic preferences across cultures, styles, and applications. Future work should investigate broader and more diverse notions of visual composition.