# Photography Perspective Composition: Towards Aesthetic Perspective Recommendation

**Lujian Yao*** **Siming Zheng†** **Xinbin Yuan** **Zhuoxuan Cai** **Pu Wu**
**Jinwei Chen** **Bo Li** **Peng-Tao Jiang†** ✉
vivo Mobile Communication Co., Ltd
lujianyao@mail.ecust.edu.cn, pt.jiang@vivo.com
Project page: https://vivocameraresearch.github.io/ppc

## Abstract

Traditional photography composition approaches are dominated by *2D* cropping-based methods. However, these methods fall short when scenes contain poorly arranged subjects. Professional photographers often employ perspective adjustment as a form of *3D recomposition*, modifying the projected 2D relationships between subjects while maintaining their actual spatial positions to achieve better compositional balance. Inspired by this artistic practice, we propose *photography perspective composition* (PPC), extending beyond traditional cropping-based methods. However, implementing the PPC faces significant challenges: the scarcity of perspective transformation datasets and undefined assessment criteria for perspective quality. To address these challenges, we present three key contributions: (1) An automated framework for building PPC datasets through expert photographs. (2) A video generation approach that demonstrates the transformation process from less favorable to aesthetically enhanced perspectives. (3) A perspective quality assessment (PQA) model constructed based on human performance. Our approach is concise and requires no additional prompt instructions or camera trajectories, helping and guiding ordinary users to enhance their composition skills.

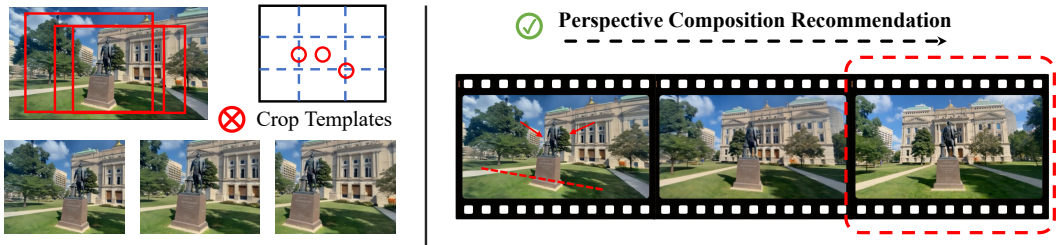
## 1 Introduction

Professional photography demands expertise in multiple aspects, with photographic composition being one of the most crucial. Photographic composition refers to the arrangement of visual elements according to aesthetic principles. It requires photographers to harmoniously integrate multiple elements like people, urban, and natural features. Master photographers, such as those in Magnum Photos, require professional knowledge and extensive training, making quality photography expensive and challenging for ordinary people. This raises the question: Can we help ordinary people achieve professional-level composition?

Traditional photography composition approaches are primarily based on cropping. Numerous approaches have been developed for image cropping, including saliency-based methods [42], learning-based techniques [6, 10, 13, 22, 28, 29, 49, 60, 62], and reinforcement learning strategies [21]. However, crop-based methods are inherently limited, as they only allow for 2D recomposition within the image plane. As shown in Fig. 1a, traditional crop-based methods primarily focus on learning a crop template. However, when the scene itself is chaotic and lacks good compositional structure, cropping alone rarely produces satisfactory results.

*Intern at vivo Mobile Communication Co., Ltd.
†Project lead.
✉Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).




(a) Crop-based Photo Composition

**(b) Our Perspective-based Photo Composition**

Figure 1: The motivation for the proposed photography perspective composition (PPC). Traditional crop-based methods (a) focus on learning crop templates for better composition. However, when scenes contain chaotic arrangements of subjects, cropping alone rarely yields satisfactory results. Perspective transformation (b) addresses these challenges by adjusting spatial relationships between subjects (e.g., person and tree, red arrow) and scene orientation.

In real-world scenarios, photographers address these limitations of 2D cropping by actively adjusting their perspective and positions to achieve improved spatial relationships between subjects. Through perspective adjustments, photographers can create sophisticated compositions by systematically arranging subjects within the frame, manipulating spatial relationships to create dynamic and engaging images. Fig. 1b illustrates how perspective transformation can address compositional challenges.

Inspired by this artistic practice, we introduce *photography perspective composition* (PPC) as a new paradigm for photography composition. However, implementing PPC presents three main challenges: ❶ Data acquisition is particularly challenging as currently available datasets are limited to planar crop data and lack perspective transformation information. ❷ The implementation of perspective recommendation requires careful design considerations, as compositional aesthetics often follow partial ordering rather than total ordering relationships [39]. ❸ The aesthetic evaluation of different perspectives requires new metrics and evaluation methods.

To address these challenges, we propose three key solutions. First, we develop a novel method for constructing aesthetic perspective transformation datasets, with an *automated data generation pipeline* (for ❶). Second, we implement a perspective transformation video generation approach instead of single-image recommendations. This enables before-and-after compositional comparisons while providing users with intuitive visual guidance (for ❷). Finally, we construct a comprehensive perspective quality assessment (PQA) model that evaluates perspective transformation quality through three critical dimensions: visual quality, motion quality, and composition aesthetic (for ❸).

Our main contributions can be summarized as follows:

- To the best of our knowledge, we are the first to introduce *photography perspective composition*, moving beyond traditional cropping methods. We hope our work can inspire more researchers to explore and advance perspective-based composition techniques in computational photography.
- We develop a concise PPC, without requiring additional prompt instructions or camera motion trajectories, which can help ordinary users improve their photo composition skills.
- We present an automated framework for constructing aesthetic perspective transformation datasets, which leverages large-scale expert photograph collections to learn generalizable principles of aesthetic composition.
- A perspective quality assessment (PQA) model is constructed based on human performance to evaluate the quality of perspective transformation through three aspects: visual quality, motion quality, and compositional aesthetics.

## 2 Related Work

This section reviews two key research areas related to our work: (1) Photography composition, focusing on various image cropping techniques and their data-driven approaches (Sec. 2.1), and (2)

the evaluation with human performance, particularly the recent adoption of vision-language models (VLM) for quality assessment (Sec. 2.2).

### 2.1 Photography Composition

Prior photography composition methods primarily rely on image cropping. There are three types. (1) free-form cropping. Numerous techniques have been explored to tackle this problem, including saliency maps [42], learning-based methods [6, 10, 13, 22, 28, 29, 42, 49, 60, 62], and reinforcement learning [21]. (2) Subject-aware image cropping [14, 55], where an additional subject mask is provided to indicate the subject of interest. (3) Ratio-aware cropping [5], where the crops are expected to adhere to a specified aspect ratio. Several datasets have been established, including GAIC [61], CPC [50], FCDB [4], and SACD [56]. Recent advances in diffusion models [36, 65] have further expanded the possibilities, with works leveraging Stable Diffusion for synthetic data generation [38, 41] and developments in outpainting [14, 40, 58]. However, cropping methods share a fundamental limitation as they operate only in 2D space, making them insufficient when the spatial arrangement of the scene is less favorable.

### 2.2 Evaluation with Human Performance

Evaluation models are essential for aligning generative models with human preferences. Traditional metrics like FID [12] and CLIP scores [33] have been widely used [16, 17, 27]. To improve evaluation accuracy, recent works [9, 18, 24, 51, 54, 64] have evolved from simple metrics to learning-based approaches that leverage human preference datasets to train CLIP-based models. Recently, researchers have begun exploring vision-language models (VLMs) [2, 47] as a more powerful framework for reward modeling. These VLM-based approaches have shown success in both evaluation [11, 51, 52] and optimization [20, 32, 45, 54], utilizing methods like point-wise regression [11, 53], pair-wise comparison with Bradley-Terry loss [3], and instruction tuning [23, 25, 48] to leverage reasoning capabilities for VLMs. However, VLMs need substantial data for effective training, yet expert compositional data is scarce and costly to obtain, making it challenging to train VLMs to capture sophisticated aesthetic principles using only limited expert annotations.

## 3 Methodology

### 3.1 Overview

Unlike previous photography composition methods that primarily rely on cropping, we propose a novel approach that recommends photography composition through *perspective transformation*. First, we describe how to construct the dataset and implement an automated pipeline (Sec. 3.2). Then, we present the core implementation methods of our PPC, and incorporate RLHF to make the generated videos aligned with human performance (Sec. 3.3). Finally, we introduce the implementation of the perspective quality assessment (PQA) Model (Sec. 3.4).

### 3.2 Automated Construction of PPC Dataset

**[Intuition.]** Currently, no dedicated dataset exists for PPC setting. To promote this area, we propose a novel approach for constructing PPC dataset in an automated way. The main challenge lies in collecting real-world camera movement sequences that transition from *less favorable* to *well-composed* perspective. The closest existing data comes from photographers sharing point-of-view (POV) recordings of their shooting process on streaming platforms. However, these videos are typically captured from secondary angles using GoPro, which differs from the main camera perspective. We observe that expertly composed photographs are readily available [43], and recent advances in single-image scene reconstruction have shown remarkable results [59]. This inspired us to explore an alternative approach: *generating camera movement sequences that transition from well-composed to less favorable perspective based on existing expertly composed photographs*. By *reversing* these sequences, we can obtain the desired data.

**[Detail.] (1) Data Source.** We select multiple professional photography datasets, including datasets used in existing composition studies such as GAIC [62], SACD [55], FLMS [8], and FCDB [4].

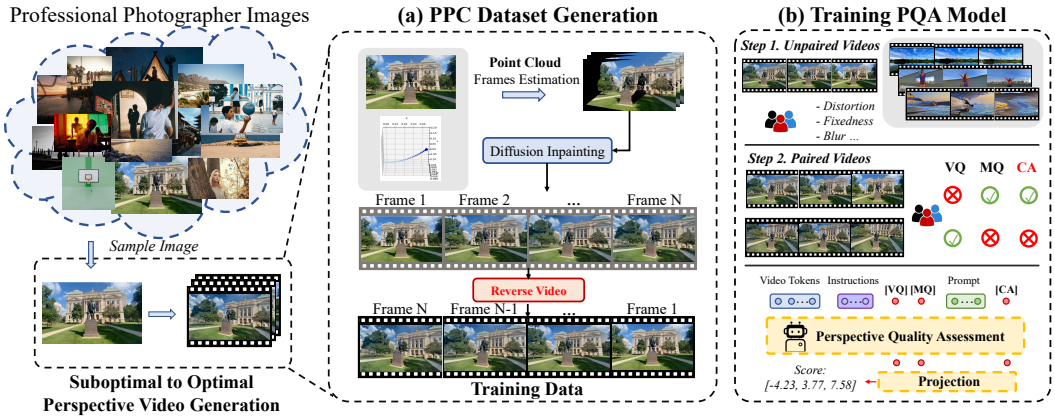



Figure 2: Architecture illustration of PPC dataset generation and the training perspective quality assessment (PQA) model.

Table 1: Grade assessment in data filtering.

| Grade | Level | Raw Score Range | Standardized Score (Range) | Quality Assessment |
|---|---|---|---|---|
| A | Excellent | $\geq 5.0$ | 95 (90-100) | Superior quality |
| B | Good | 0.0 to 4.9 | 85 (80-89) | Above-average quality |
| C | Satisfactory | -5.0 to -0.1 | 75 (70-79) | Acceptable quality |
| D | Marginal | -15.0 to -5.1 | 65 (60-69) | Below-average quality |
| E | Unsatisfactory | $< -15.0$ | 50 ($< 60$) | Inadequate quality |

Furthermore, to expand our data volume, we incorporated Unsplash [1], currently the largest open-source professional photography dataset. **(2) Perspective Transformation Generation.** As shown in Fig. 2, we adopt a 3D reconstruction approach. Our 3D reconstruction methodology mainly builds upon the ViewCrafter [59]. The inputs consist of a well-composed image and a specified camera motion trajectory. Note that this trajectory can be random (refer to **[Discussion]**). By following this trajectory, we can generate a video sequence transitioning from the well-composed to less favorable perspective. Then, by *reversing* this video sequence, we obtain our desired training data. **(3) Data Filtering.** Given the limited performance of current reconstruction models, the generated video data needs to filter out artifacts including distortion, fixedness, and blur effects, as depicted in Fig. 2. However, manual filtering for such a large dataset is impractical. Our tests showed that a single person can only filter about 3K videos per day, making it difficult to process large-scale samples. With the rapid advancement of vision language models (VLMs) in scene understanding and automated evaluation [26, 32, 52], we develop a perspective quality assessment (PQA) model to filter the generated data. For specific details about the PQA construction, please refer to Sec. 3.4. Our PQA evaluates generated data across three dimensions: visual quality (VQ), motion quality (MQ), and composition aesthetic (CA). These individual scores are aggregated into a final score, and samples exceeding a threshold are selected as our training data. We implement a comprehensive grading assessment that converts model-generated scores into standardized grade scores. It employs a five-tier grading scale (A to E) with corresponding numerical ranges, as shown in Tab. 1.

**[Discussion.]** *How to handle cases where the initial perspective is less favorable?* In our pipeline, we initially treat the original image as the well-composed perspective. This assumption is later refined through human preference learning during the PQA filtering (Sec. 3.4) and RLHF (Sec. 3.3) stages, acknowledging that the original perspective may not always be the aesthetically pleasing.

### 3.3 Photography Perspective Composition (PPC)

**[Intuition.]** Previous image composition works primarily focused on cropping approaches. In contrast, we propose a video-based approach: given a less favorable perspective, we generate a camera movement sequence that gradually transitions to an aesthetically enhanced perspective of the scene. This video-based approach is motivated by two key observations: First, image composition represents

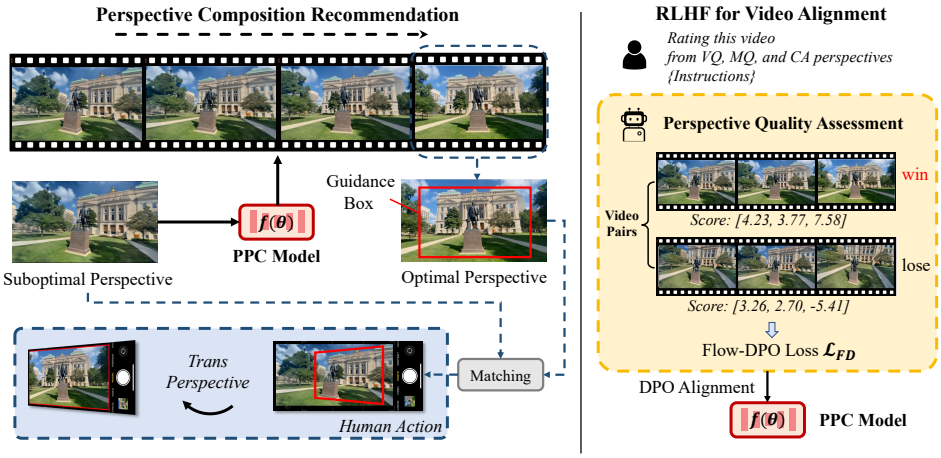



Figure 3: The pipeline of proposed photography perspective composition (PPC).

a partial order relationship where quality assessment often relies on comparing different views of the same scene rather than isolated evaluation. Second, the transition process naturally demonstrates compositional improvements through before-and-after comparisons, making it particularly effective for both visual demonstration and educational purposes.

**[Detail.] (1) Base Pipeline.** As shown in Fig. 3, our pipeline takes a less favorable perspective as input and generates a transformation video from the less favorable to aesthetically enhanced perspective. This process can be modeled as an image-to-video (I2V) task. I2V has seen remarkable progress, with both commercial solutions like OpenAI's Sora [30], Runway Gen-3 [37], and Pika 1.5 [31], and open-source models like Hunyuan [19], CogVideo [57], and WAN [46]. Our pipeline leverages these open-source models to generate perspective transformation videos. We utilize the last frame of the video as our final aesthetically enhanced perspective and design a method to guide human actions. First, we draw a guidance box (the red bbox in Fig. 3) on the enhanced perspective. Then, based on this box, along with the initial and final perspectives, we transform this box onto the original image using feature matching, creating a distorted box. As the user moves, this box gradually changes shape, approaching a rectangle when reaching the aesthetically enhanced perspective. To simplify the process and accelerate computation, we only use traditional *homography* transformation [7].

**(2) RLHF for Quality Enhancement.** We observed that some generated perspective transformation videos, while deviating from the GT direction, still maintain high aesthetic quality. This suggests that strict GT adherence may not always yield the most aesthetically pleasing results. Therefore, we propose incorporating direct preference optimization (DPO) to align the model with human preferences. This approach encourages the exploration of aesthetically pleasing trajectories that may differ from GT, avoiding the limitation where GT-based optimization could discourage potentially superior compositional alternatives.

Our RLHF implementation primarily draws from Diffusion-DPO [34] and VideoAlign [26]. Consider a fixed dataset $\mathcal{D} = \{s, v_h, v_l\}$, where each sample consists of a prompt $s$ and two videos, $v_h$ (higher-quality video) and $v_l$ (lower-quality video), generated by a reference model $p_{\text{ref}}$, with human annotations indicating that $v_h$ is preferred over $v_l$ (i.e., $v_h \succ v_l$). The goal of RLHF is to learn a conditional distribution $p_\theta(v \mid s)$ that maximizes a reward model $r(v, s)$ (i.e., PQA model proposed in Sec. 3.4), while controlling the regularization term (KL-divergence) from the reference model $p_{\text{ref}}$ via a coefficient $\beta$:

$$\max_{p_\theta} \mathbb{E}_{s\sim\mathcal{D}_c, v\sim p_\theta(v|s)} \left[r(v, s)\right] - \beta\, \mathbb{D}_{\text{KL}} \left[p_\theta(v \mid s) \,\|\, p_{\text{ref}}(v \mid s)\right]. \tag{1}$$

In Rectified Flow, we adopt videoalign [26] that relate the noise vector $\xi^*$ to a velocity field $\nu^*$, where $\nu^*$ represents the velocity field of either the higher-quality video ($\nu^h$) or lower-quality video ($\nu^l$). Specifically, it can be proved that $\|\xi^* - \xi_{\text{pred}}(\nu_t^*, t)\|^2 = (1-t)^2 \|\nu^* - \nu_{\text{pred}}(\nu_t^*, t)\|^2$ , where $\xi_{\text{pred}}$ and $\nu_{\text{pred}}$ refer to predictions either from the model $p_\theta$ or the reference model $p_{\text{ref}}$. Based on this

Table 2: Comparison of I2V models in generating perspective transformation videos.

| Method | Perspective Accuracy | | | | | Human Performance Score | | |
|---|---|---|---|---|---|---|---|---|
| | CMM ↑ | FVD ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | VQ ↑ | MQ ↑ | CA ↑ |
| CogvideoX 1.5 5B [57] | 0.5501 | 303 | 8.2380 | 0.2611 | 0.7969 | 0.7073 | 0.7311 | **0.7196** |
| Hunyuan I2V [19] | 0.4928 | **264** | **9.4017** | **0.3537** | 0.7915 | **0.7216** | **0.7496** | 0.7070 |
| Wan2.1 14B [46] | **0.5989** | 345 | 9.3668 | 0.3265 | **0.7808** | 0.7195 | 0.7454 | 0.7072 |

| Method | Video Quality | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | I2V Subject | I2V Background | Subject Consistency | Background Consistency | Motion Smoothness | Dynamic Degree | Aesthetic Quality | Imaging Quality |
| CogvideoX 1.5 5B [57] | 0.9632 | 0.9639 | 0.9545 | 0.9502 | 0.9906 | 0.1347 | 0.5314 | 0.5667 |
| Hunyuan I2V [19] | **0.9866** | **0.9878** | **0.9582** | **0.9663** | **0.9927** | 0.7851 | **0.5583** | 0.5893 |
| Wan2.1 14B [46] | 0.9618 | 0.9694 | 0.9470 | 0.9435 | 0.9917 | **0.8883** | 0.5464 | **0.6299** |

relationship, we obtain the Flow-DPO loss $\mathcal{L}_{\text{FD}}(\theta)$:

$$-\mathbb{E}\left[\log\sigma\left(-\frac{\beta_t}{2}\left((\|\nu^h-\nu_\theta(\nu_t^h,t)\|^2-\|\nu^h-\nu_{\text{ref}}(\nu_t^h,t)\|^2)-\left(\|\nu^l-\nu_\theta(\nu_t^l,t)\|^2-\|\nu^l-\nu_{\text{ref}}(\nu_t^l,t)\|^2\right)\right)\right)\right], \tag{2}$$

where $\beta_t = \beta\,(1-t)^2$ and the expectation is taken over samples $\{v_h, v_l\} \sim \mathcal{D}$ and the schedule $t$.

### 3.4 Perspective Quality Assessment (PQA) Model

**[Intuition.]** PQA serves dual purposes: filtering generated training data to enable automated dataset construction, and providing win-lose pairs for RLHF in PPC to align with human preferences. Due to the limitations of 3D reconstruction models, some scenes suffer from inaccurate point cloud reconstruction or inpainting distortions, leading to distortion, fixedness, and blur. We propose a *two-stage* training strategy for the PQA model that addresses the dual challenges of data volume and expertise requirements. Given that fine-tuning VLMs demands substantial training data, the first *unpair-wise* stage leverages large-scale, efficiently collected data to meet this requirement, as basic quality assessment does not demand expert knowledge. The subsequent *pair-wise* stage employs expert-annotated compositional data to refine the aesthetic capabilities of the model.

**[Detail.] (1) Dateset Setting.** *Stage ①: Unpaired Videos.* This stage focuses on distinguishing video quality levels. We collected approximately 5K perspective transformation videos generated by 3D reconstruction models, with expert annotators identifying roughly 1.5K high-quality and 3.5K low-quality samples. To expand the dataset, we randomly paired each high-quality video with 10 low-quality ones, creating a 15K unpaired dataset. *Stage ②: Paired Videos.* This stage focuses on composition aesthetic recognition through paired comparison learning. For each initial perspective, we generate three video clips using separately trained CogVideoX 1.5, WAN 2.1, and the original GT data. These clips are paired with each other, where "paired" indicates videos sharing identical input views. Expert annotators evaluate these pairs across three dimensions: visual quality (VQ), motion quality (MQ), and composition aesthetic (CA). For each dimension, annotators choose between options (*A wins/Ties/B wins*). Notably, the CA metric assesses the compositional improvement throughout the video transformation rather than static frame quality. Detailed annotation guidelines are provided in Appendix A.

**(2) Model Setting.** Our model primarily follows the architecture of VideoAlign [26]. We employ Qwen 2-VL as our base model, utilizing the Bradley-Terry model with ties loss (BTT) [35], which extends the traditional Bradley-Terry framework[3]. To better handle multi-dimensional evaluation, we separate special tokens for context-agnostic (visual quality, motion quality) and composition-aware (composition aesthetic) attributes, leveraging the causal attention mechanism to achieve effective feature decoupling. The model predicts rewards for each dimension through a shared linear projection head applied to the corresponding token representations from the final layer.

## 4 Experiments

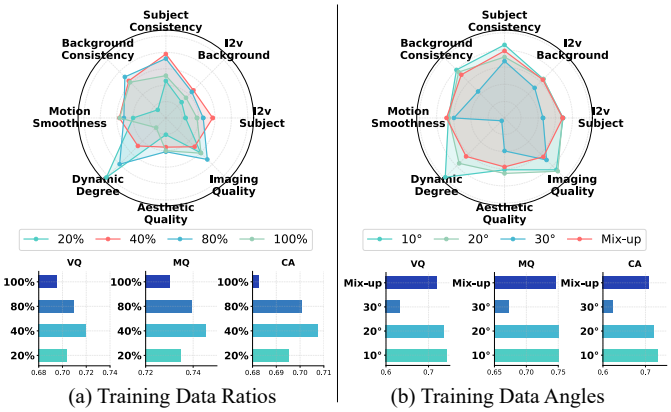
In this section, we evaluate our approach through two main components: photography perspective composition (PPC) and perspective quality assessment (PQA).

Figure 5: PPC performance in single-subject scenarios.

### 4.1 Investigation for Photography Perspective Composition (PPC)

**Implementation.** To comprehensively validate our approach, we experimented with three state-of-the-art video generation models: CogVideoX 1.5 [57], HunYuan [19], and Wan2.1 [46]. The training parameters follow the settings from the original repository.

**Metric Design.** Since there is no prior work on photography composition using perspective transformation, we need to define metrics. We evaluate the generated videos from three parts: perspective accuracy, video quality, and human performance score. For *video quality* assessment, we primarily adopt the evaluation metrics from VBench2.0 [17] I2V benchmarks, which include I2V subject, I2V background subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and image quality. For *perspective accuracy* evaluation, we employed both video distance and image similarity metrics. The video distance was measured using camera motion matching (CMM) and Fréchet Video Distance (FVD) [44] , while image similarity was assessed using CLIP-F [33], PSNR, SSIM, and LPIPS. Notably, we modified the camera motion section of the original vbench [17], changing it to camera motion matching. As we found that the camera motion detection model in vbench did not work well, especially for small angle changes. Instead, we now pass both the prediction and ground truth through the camera motion detection and then calculate the accuracy by matching these two detection results." As for the *human performance score*, its purpose is to simulate human preferences in scoring. This is conducted through PQA, which we established in Sec. 3.4, and includes three components: visual quality (VQ), motion quality (MQ), and composition aesthetic (CA).




(a) Training Data Ratios (b) Training Data Angles

Figure 4: Quality and human performance results for PPC.

**Main Results.** *[Quantitative Results]* As shown in Tab. 2, we demonstrate the performance of three I2V models in generating PPC videos. *[Qualitative Results]* We demonstrate the versatility of our approach across three representative scenarios. The first scenario addresses *single subjects* (e.g., human figures and animals), as shown in Fig. 5, where our PPC model enhances compositional harmony by seamlessly integrating subjects with their surroundings. In the second scenario (Fig. 7), we tackle *multi-subject* scenes, demonstrating how PPC achieves balanced spatial arrangements

Table 3: Quantitative Result for PQA (Tab. (a) and Tab.(b)) and PPC (Tab. (c) and Tab. (d)).

(a) The number of pairs.

| # Pairs | Accuracy | | |
|---|---|---|---|
| | VQ | MQ | CA |
| 1 | 0.6515 | 0.5957 | 0.5881 |
| 5 | 0.7772 | 0.7812 | 0.7894 |
| 10 | 0.8019 | 0.8085 | 0.8102 |
| 100 | 0.8008 | 0.8063 | 0.8103 |

(b) Different steps.

| # Steps | Accuracy | | |
|---|---|---|---|
| | VQ | MQ | CA |
| Reg Single | 0.4874 | 0.4986 | 0.4761 |
| Reg Two | 0.7807 | 0.7802 | 0.7935 |
| BTT Single | 0.5509 | 0.5117 | 0.4913 |
| BTT Two | 0.8019 | 0.8085 | 0.8102 |

(c) The data ratios.

| Ratio | CMM ↑ | FVD ↓ |
|---|---|---|
| 20% | 0.5014 | 460 |
| 40% | 0.5989 | 345 |
| 80% | 0.5244 | 362 |
| 100% | 0.5673 | 359 |

(d) The data angles.

| | CMM ↑ | FVD ↓ |
|---|---|---|
| 10° | 0.4413 | 397 |
| 20° | 0.5587 | 337 |
| 30° | 0.3983 | 444 |
| Mix-up | 0.5989 | 345 |

to elevate overall visual aesthetics. The third scenario explores *landscape* photography (Fig. 8), addressing two common challenges faced by novice photographers: balance and horizontal alignment. Our model effectively optimizes these scenes, particularly enhancing symmetrical compositions. Beyond these scenarios, we discovered the applicability of PPC to UAV photography. As illustrated in Fig. 9, our model successfully identifies aesthetically enhanced views from drone-like perspectives, generating camera movements that adhere to compositional principles while maintaining aesthetic appeal.

**PPC Angles.** To maintain high consistency in our recommendation system, we limit perspective transformations to short angles, ensuring our suggestions are reliably based on the actual visible content in the current scene. We also investigated the impact of PPC angles on performance using three distinct rotation degrees (10°, 20°, and 30°) and a balanced mixed dataset incorporating all angles. As shown in Tab. 3d and Fig. 4b, we observe that while performance remains stable at 10°, both quality and accuracy metrics deteriorate significantly when the dataset rotation angles reach 30°. We attribute this degradation to the substantial visual disparity between the original and transformed views at larger angles, which makes it challenging for the model to learn generalized aesthetic perspectives.

**The Consistency of PPC.** Fig. 6b demonstrates our model's consistency in PPC. When presented with different less favorable views of the same scene, our model generates consistent aesthetically enhanced perspectives, maintaining coherence across different inputs.

**The Effection of RLHF in Video Quality Enhancement.** We evaluate the performance after incorporating RLHF (Sec. 3.3). Fig. 6a and Tab. 4 demonstrate the effectiveness of incorporating RLHF. The results show that RLHF leads to more stable subject generation.


(a) Qualitative comparison of RLHF

(b) PPC maintains perspective consistency

Figure 6: Qualitative comparison of RLHF and the perspective consistency of PPC.

### 4.2 Investigation for Perspective Quality Assessment (PQA)

**Implementation and Evaluation Metric.** We utilize Qwen2-VL-2B [47] as the backbone for PQA and train it with BTT loss. To fine-tune the model, LoRA [15] is applied to update all linear layers in the language model, while the vision encoder's parameters are fully optimized. The training process is conducted with a batch of 32 and a learning rate of $2 \times 10^{-6}$, with the model trained over two epochs. This setup requires approximately 50 NVIDIA H20 GPU hours. Several observations were made during training. We sample videos at 1 fps, with a resolution of $448 \times 448$ pixels during the training process. Following previous works [26], we adopt accuracy as the metric for each dimension.

**Main Results.** *[Quantitative Result]* To investigate the impact of training data volume in *stage 1*, we conducted experiments with varying numbers of video pairs, as detailed in Tab 3a. Our results reveal that while performance generally improves with increasing sample size, it plateaus at approximately 100 samples, suggesting that this quantity provides sufficient diversity for model performance. *[Qualitative Results]* Fig. 10 demonstrates the basic effects of PQA. As shown in

Table 4: The effect of RLHF in PPC.

| | CMM | FVD | VQ | MQ | CA |
|---|---|---|---|---|---|
| w/o RLHF | 0.4928 | 264.7672 | 0.7216 | 0.7496 | 0.7070 |
| with RLHF | 0.5014 | 270.2212 | 0.7477 | 0.7774 | 0.7342 |

Figure 7: PPC performance in multiple-subject scenarios.

Figure 8: PPC performance in wide landscape views and asymmetric scenarios.

Fig. 4a, while there is a notable improvement in composition, the video quality remains low. Fig. 4b exhibits acceptable levels of both VQ and MQ, but shows minimal compositional differences, resulting in a lower CA score. Notably, PQA assigns lower ratings to cases where the perspective remains static or too intense, as illustrated in the last two examples.

**Effect of Two Steps.** As shown in Tab.3b, we evaluated the effectiveness of the two-step approach under both regression and BTT loss [3, 26] functions. The single-step approach, lacking sufficient data to enhance baseline performance, demonstrates significantly lower overall performance compared to the two-step way.

## 5 Conclusion and Limitation

**Conclusion.** In this work, we addressed the limitations of previous photography composition methods by introducing photography perspective composition (PPC), a novel paradigm that extends beyond 2D cropping to achieve 3D recomposition. Our approach is inspired by real-world street photography practices where photographers use perspective adjustment to establish better relative relationships between subjects. To overcome the challenges of implementing PPC, particularly the lack of suitable datasets and unclear assessment criteria, we made three significant contributions. We developed a framework for automatically constructing a PPC dataset from expert photographs, created a system for generating perspective transformation videos that guide users from less favorable to aesthetically enhanced views, and introduced the perspective quality assessment (PQA) model that evaluates both video quality and compositional aesthetics. We hope this work opens up new possibilities in computational photography and inspires further research in perspective-aware composition.

Figure 9: PPC performance in UAV-like scenarios.


Figure 10: Quantitative Results of PQA.

**Limitation and Future Work.**

**(1) Video Duration.** Current PPC is constrained by the limitations of existing video models, particularly in terms of duration. In the latest phase, we have also discovered AR-based video generation models that can generate infinite streaming videos. We believe this work can provide broader insights and directions for PPC.

**(2) Video Quality.** Since our training data is generated by 3D reconstruction models, the performance of PPC is inherently limited by current reconstruction capabilities. A crucial direction for future improvement lies in exploring superior methods for generating perspective transformation videos, such as utilizing the *Unreal Engine 5* for video generation.

**(3) Data Scaling Behavior.** Despite the strong scene diversity provided by numerous expert photography images, we observed that model outputs become unstable as the training data volume increases. As shown in Tab. 3c, Fig. 4a, and Fig. 11, while both accuracy and quality initially improve with increasing data volume, performance deteriorates when the training set size grows further. We hypothesize that this challenge lies in maintaining the desired model behavior to ensure proper perspective rather than deviating into unintended random behaviors. This phenomenon bears similarities to what was described in IC-Light [63], and we plan to incorporate their training methodology to investigate whether it can enhance training stability in our future work.

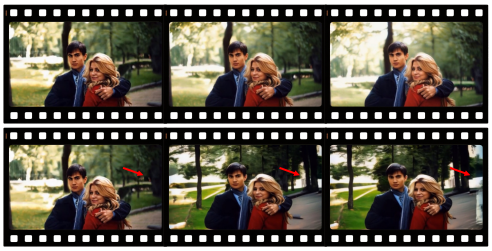


Figure 11: Diverse data presents instability.

## References


[1] Unsplash dataset. https://unsplash.com/, 2023.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[4] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2017.

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

[6] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 870–878, 2018.

[7] Elan Dubrofsky. Homography estimation. *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, 5, 2009.

[8] 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 ACM International Conference on Multimedia*, 2014.

[9] Michael Fischer, Konstantin Kobs, and Andreas Hotho. Nicer: Aesthetic image enhancement with humans in the loop. *arXiv preprint arXiv:2012.01778*, 2020.

[10] 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 Transactions on Multimedia*, 20(8):2073–2085, 2018.

[11] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024.

[12] 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. *Advances in Neural Information Processing Systems*, 30, 2017.

[13] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7057–7066, 2021.

[14] James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning subject-aware cropping by outpainting professional photos. *arXiv preprint arXiv:2312.12080*, 2023.

[15] 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.

[16] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[17] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[18] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

[19] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[20] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[21] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8193–8201, 2018.

[22] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4213–4222, 2020.

[23] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*, 2023.

[24] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411, 2024.

[25] Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. Criticbench: Benchmarking llms for critique-correct reasoning. *arXiv preprint arXiv:2402.14809*, 2024.

[26] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

[27] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024.

[28] Weirui Lu, Xiaofen Xing, Bolun Cai, and Xiangmin Xu. Listwise view ranking for image cropping. *IEEE Access*, 7:91904–91911, 2019.

[29] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 497–506, 2016.

[30] OpenAI. Video generation models as world simulators, 2024.

[31] PikaLabs. Pika 1.5. *https://pika.art/*, 2024.10.

[32] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[34] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[35] PV Rao and Lawrence L Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022.

[37] Runway. Gen-3. *https://runwayml.com/*, 2024.06.

[38] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023.

[39] Herbert A Simon and Allen Newell. Human problem solving: The state of the theory in 1970. *American psychologist*, 26(2):145, 1971.

[40] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T. Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2019.

[41] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners, 2023.

[42] 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, pages 12104–12111, 2020.

[43] Unsplash. Unsplash dataset, 2023.

[44] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. *Openreview*, 2019.

[45] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[46] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[47] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[48] Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024.

[49] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2018.

[50] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[51] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[52] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.

[53] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

[54] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[55] 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. *Computational Visual Media*, 9(1):87–107, 2023.

[56] 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. *Computational Visual Media*, 9(1):87–107, 2023.

[57] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[58] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2019.

[59] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

[60] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5949–5957, 2019.

[61] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.

[62] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1304–1319, 2020.

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025.

[64] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2024.

[65] Tianyi Zheng, Peng-Tao Jiang, Ben Wan, Hao Zhang, Jinwei Chen, Jia Wang, and Bo Li. Beta-tuned timestep diffusion model. In *European Conference on Computer Vision*, pages 114–130. Springer, 2024.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations in the Conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper has no theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information necessary to reproduce the main experimental results (in main submission and appendix), ensuring that the main claims and conclusions of the paper can be independently verified and validated.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper has not yet been open-sourced for data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide a thorough description of the details of our experiments in the paper and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: There is no reporting of error bars or statistical significance information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: : The paper does not provide sufficient information on the computer resources, such as the type of compute workers, memory, and time of execution needed to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms to the NeurIPS Code of Ethics, as outlined in the provided URL. The paper adheres to the ethical practices and guidelines specified in the NeurIPS Code of Ethics during the research process.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper solely emphasizes the positive societal impacts of the work performed, omitting any discussion of potential negative consequences or societal drawbacks.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper thoroughly acknowledges and properly credits the creators or original owners of assets, including code, data, and models, used in the research. Additionally, it explicitly mentions and respects the licenses and terms of use associated with these assets, ensuring ethical and legal compliance.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper contributes a new dataset for PPC.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We adopt Qwen-2-vl as the base model in constructing PQA.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

# Appendix

## A Prompt Instructions

For VQ and MQ, our instructions primarily follow those of VideoAlign [26]. We specifically designed the CA instructions as shown below:

**Input Template (CA part)**

****Composition Aesthetic**:**
Evaluate the evolution and sophistication of compositional techniques throughout the video, drawing inspiration from Magnum Photos' aesthetic principles. Consider the following sub-dimensions:
- **Layering Complexity**: Assess how the video utilizes multiple planes and creates depth through foreground, middle ground, and background interactions. Consider if these spatial relationships become more sophisticated over time.
- **Geometric Harmony**: Evaluate the use of strong geometric elements, lines, and shapes that create dynamic tension and visual interest, similar to Alex Webb's approach to complex frame organization.
- Color Relationships: Consider how color blocks and contrast are used compositionally to create visual weight and guide viewer attention through the frame.
- **Frame Utilization**: Evaluate how effectively the entire frame is used, including edges and corners, and how secondary elements support the main subject.
- **Visual Rhythm**: Consider the pattern and repetition of elements, and how they create compositional flow and movement within the frame.
- **Juxtaposition Development**: Assess how the video develops and maintains meaningful visual relationships between different elements in the frame.

Please provide the ratings of Composition Aesthetic: `<|CA_reward|>`
END

**Instruction Source.** The compositional principles in our framework are derived from seminal works in photography theory and practice. These include the layering complexity theory from Alex Webb's "The Suffering of Light" and Sam Abell's three-layer composition approach; geometric harmony principles from Henri Cartier-Bresson's "The Decisive Moment"; color relationship theories from Steve McCurry and Ernst Haas; frame utilization techniques from Robert Frank's "The Americans"; and visual rhythm concepts from Paul Strand and Minor White. These principles, extensively documented in the works of Magnum Photos photographers, form the theoretical foundation for our composition assessment criteria.

Table 5: Key points summary outlined in annotation guidelines for CA evaluation dimension.

| Evaluation Dimension | Key Points Summary |
|---|---|
| Composition Aesthetic | Considering the following dimensions in the compositional design of the video:<br>- **Compositional Reasonableness**: The composition should be objectively reasonable and well-balanced.<br>- **Compositional Clarity**: The arrangement of elements should be clear and visually organized.<br>- **Compositional Detail**: The level of sophistication in the arrangement and relationship between elements.<br>- **Compositional Creativity**: The composition should be aesthetically pleasing and show creative arrangement.<br>- **Compositional Safety**: The composition should not create visual tension or uncomfortable viewing experience. |