ViewRec3D: Learning to Recommend 3D Viewpoints for AI Photography

Chengzhi Zhang[1,*] **Yunzhong Hou**[1,*,†] **Zizheng Sun**[1] **Ming-Hsuan Yang**[2] **Chi Harold Liu**[1]
[1]Beijing Institute of Technology [2]University of California, Merced
[*]Equal contribution. [†]Corresponding author.

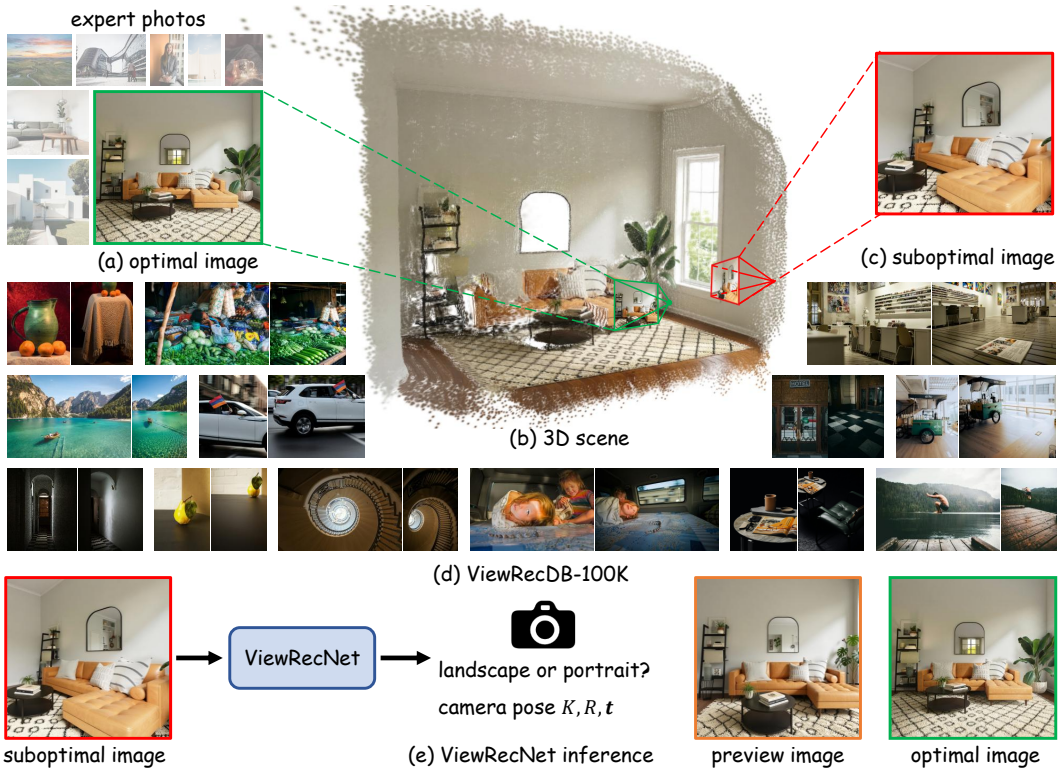


Figure 1: Treating an expert photo as the optimal viewpoint (a), we first outpaint the image and reconstruct the 3D scene (b), and then render random alternative viewpoints and select a high-quality result as the suboptimal image (c). In this manner, we build data pairs for viewpoint recommendation in AI photography, and curate the ViewRecDB-100K dataset (d). Using this dataset, we train a viewpoint recommendation model called ViewRecNet. Given an image as input, ViewRecNet predicts the expert viewpoint (e), which mimics that of the optimal image.

Abstract

Viewpoint dictates the photo composition and plays a vital role in AI photography. Existing 3D viewpoint recommendation methods are often *limited* in viewpoint adjustments, which is mainly due to their lack of high-quality training data with *large and accurate* 3D view changes. Ideally, the training data should contain suboptimal images and expert-selected optimal viewpoints, but this can be very expensive to curate. To solve this, we build an automatically generated 3D viewpoint recommendation dataset from expert photos. Specifically, we first regard these expert photos

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

as optimal and outpaint them to provide more information on scene layouts. Next, we explicitly reconstruct the 3D scene. We then render suboptimal images from random viewpoints. Lastly, we introduce a hierarchical data filtering scheme and acquire high-quality data pairs for our ViewRecDB-100K dataset. On top of this, we further introduce a viewpoint recommendation model, ViewRecNet, that can predict expert viewpoints from any suboptimal image inputs using ray-based dense supervision. The proposed method achieves strong results both quantitatively and qualitatively, supported by LLM-based viewpoint preference and human studies. Data and code are available at link.

1 Introduction

Camera viewpoints play a major part in determining photo composition, which involves arranging visual elements within the frame to create a compelling, organized image that guides the viewer's eye [11]. Since photo composition can significantly affect the overall visual appeal, for AI photography systems that aim to help users capture better photos, viewpoint recommendation becomes an essential component with significant practical value.

Despite the practical importance of viewpoint recommendation in AI photography, existing methods still struggle to propose high-quality 3D viewpoints. In fact, these proposed viewpoints are often under limited changes when compared to the original suboptimal viewpoint, *e.g.*, very small or no significant translation and rotation changes [9, 28]. We believe that this limitation is mainly due to the lack of high-quality training data with large and accurate viewpoint changes.

In this work, we introduce a 3D viewpoint recommendation dataset, ViewRecDB-100K, that includes suboptimal images and expert-selected 3D viewpoints, as illustrated in Fig. 1. To avoid the high cost of manually curating such data from scratch, we introduce an automatic data generation method. We consider the expert-selected best-looking photos from the Unsplash dataset [17] as optimal images following [6, 28], and render suboptimal images from random viewpoints. Specifically, we adopt a 3D-reconstruction-based approach when rendering the suboptimal images. First, we outpaint the optimal image to expand the scene with possible layouts. Next, we reconstruct the expanded image as 3D Gaussian splats [7]. We then randomly select alternative viewpoints for image rendering and adopt an open-source model [4] for perspective restoration and hole filling. Lastly, we introduce a hierarchical filtering scheme that progressively filters out low-quality outputs across all stages, including image outpaints, 3D viewpoints, renders, and final restored images. With the help of a 3D expert model VGGT [18] and an LLM [16], the filtering pipeline collects high-quality data pairs with a wide range of viewpoint changes. Overall, we curate 104,704 data samples from diverse scenarios and construct the ViewRecDB-100K dataset.

On top of the dataset, we introduce a viewpoint recommendation model, termed ViewRecNet. Given an input image, ViewRecNet predicts an expert viewpoint represented by the view orientation and camera pose (intrinsics and extrinsics). Additionally, we convert camera pose changes from suboptimal images to optimal images into relative changes in terms of rays, and leverage ray-based dense supervision [31]. Using a novel view synthesis approach similar to that used in dataset generation, we render the recommended viewpoints as image previews (see Fig. 1).

Lastly, for evaluation, we build the benchmark using 1,000 professionally verified data samples. We also introduce an LLM-based viewpoint preference rater that uses chain-of-thought (CoT) reasoning. Experimental results on this benchmark demonstrate that ViewRecNet achieves strong performance both quantitatively and qualitatively.

2 Related Work

Photography Composition Optimization. Photography composition optimization aims to improve subject placement, scale, and spatial relationships with other visual elements. Existing methods primarily achieve this through 2D cropping. FCDB [1] provides human-annotated cropping windows and pairwise preferences, CPC [23] ranks candidate views, GAICD [30] assigns quality scores to grid-based candidate crops, and SACD [27] provides subject-aware cropping annotations and rankings. Other methods learn compositional preferences from professional photographs [2], or

Table 1: Viewpoint recommendation dataset comparison. #Scenes indicates the number of collected photos that are considered optimal. #Images denotes the total number of candidate viewpoints.

Dataset	2D/3D	Generation method	#Scenes	#Images	Open Source
FCDB [1]	2D	Expert Annotation	1,743	34,130	✓
CPC [23]	2D	Expert Annotation	10,800	259,200	✓
GAICD [30]	2D	Expert Annotation	1,236	106,296	✓
SACD [27]	2D	Expert Annotation	2,777	more than 24,000	✓
PPC [28]	3D	Video-based Generation	5,000	5,000	✗
PCARD [9]	3D	Panorama Sampling	4,000	324,000	✗
ViewRecDB-100K (Ours)	3D	3D Reconstruction	76,010	104,704	✓

construct training pairs by outpainting professional photographs, as in GenCrop [6]. Regarding model design, CACNet [5] incorporates composition patterns to guide cropping, while Venus [3] leverages multimodal large language models to provide interpretable aesthetic guidance and cropping results. However, 2D cropping is constrained by the original viewpoint and cannot alter perspective and occlusion relationships through camera movement.

Geometry-Aware Image Synthesis. Feed-forward geometry models [22, 18, 10] have made significant progress in recovering 3D structure and camera parameters from images. For single-image inputs, MoGe [21] estimates dense geometry for open-domain images, while SHARP [13] predicts 3D Gaussian representations for efficient novel view synthesis. Such explicit 3D representations support controllable rendering under specified camera transformations. Meanwhile, image editing models [25, 8, 12, 20, 26] demonstrate strong capabilities in object addition and removal, attribute modification, background replacement, and style transfer, producing high-quality, photorealistic images while preserving content unrelated to the intended edits. These advances support our data generation pipeline, which combines geometry-based novel view rendering for controllable viewpoint changes with generative refinement to complete missing regions and reduce rendering artifacts.

Viewpoint Recommendation. 3D viewpoint recommendation improves photography composition by adjusting camera position, rotation, and field of view (FoV). The types and ranges of viewpoint changes represented in the training data determine the adjustments a model can learn. SPAS [9] constructs PCARD by sampling views with different orientations from panoramic images. Since these views share the same camera center, the resulting training pairs contain no translation changes. PPC [28] uses ViewCrafter [29] to generate viewpoint transformation data from expert photos and trains a video generation model for viewpoint recommendation. However, its data generation process restricts rotation to a relatively small range to maintain visual quality and content consistency. In contrast, our pipeline constructs training pairs that cover camera rotation, translation, and FoV changes, as well as landscape/portrait framing adjustments, while preserving valid samples with large viewpoint variations. The resulting dataset therefore covers more types of viewpoint changes and a wider range of adjustments, enabling more flexible viewpoint recommendation.

3 The ViewRecDB-100K Dataset

3.1 Data Generation

Our large-scale 3D viewpoint recommendation dataset, ViewRecDB-100K, contains 76,010 optimal images and 104,704 suboptimal images. Following [6, 28], optimal images are collected from the Unsplash dataset [17], which contains 6.5M high-quality photos. We only consider "featured" photos, which include roughly 300k editor-selected high-quality images. We then generate the suboptimal images and the corresponding viewpoint changes (see Fig. 2). Each optimal image defines a unique scene and is associated with one or two suboptimal images. Compared with existing viewpoint recommendation datasets, ViewRecDB-100K contains the largest number of scenes, demonstrating its diversity (Tab. 1).

As shown in Fig. 2, given the optimal image I, we aim to generate a suboptimal image $\tilde{I}$ from a novel viewpoint via outpainting, 3D reconstruction, novel view rendering, and lastly image restoration. We record the viewpoint change from suboptimal image $\tilde{I}$ to optimal image I as:

$$\boldsymbol{v} = (\boldsymbol{t}, \boldsymbol{q}, \boldsymbol{s}) \,, \tag{1}$$

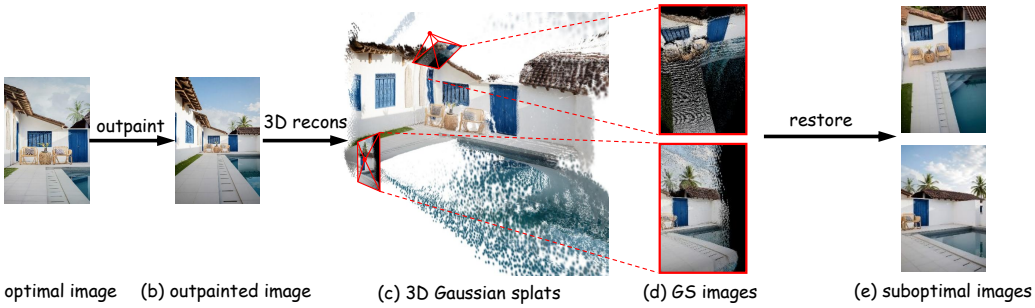


Figure 2: **Data generation method.** Given an optimal image (a), we first determine whether orientation adjustment is required and outpaint the image (b). We then reconstruct the complete 3D scene (c) and render multiple GS (Gaussian splatting) images (d) from random viewpoints. Finally, we fill holes and remove other artifacts in GS images to get corresponding suboptimal images.

where $\boldsymbol{t} \in \mathbb{R}^3$ and $\boldsymbol{q} \in \mathbb{R}^4$ denote the translation vector and rotation quaternion from $\tilde{I}$ to I, respectively, and $\boldsymbol{s} \in \mathbb{R}^2$ denotes the FoV change ratio from $\tilde{I}$ to I.

Outpainting. We use Qwen-Image-Edit [25] to outpaint I to provide a possible estimate of the scene. This step enlarges the field of view (FoV) while keeping the optimal view centered. As a result, subsequent viewpoint changes are more likely to preserve a large portion of the content in the optimal view, making the later restoration of the optimal viewpoint rely more on visible evidence instead of model priors. The entire outpainting module can be formulated as:

$$I_{\text{outpaint}} = \mathcal{F}_{\text{outpaint}}\left(I, o, \boldsymbol{s}^{-1}\right), \tag{2}$$

where $o \in \{0, 1\}$ denotes the binary orientation-change decision, and $\boldsymbol{s} \in \mathbb{R}^2$ denotes the FoV change ratio from I_{outpaint} to I. We feed the inverse of the FoV change ratio $\boldsymbol{s}^{-1}$ since the operation $\mathcal{F}_{\text{outpaint}}(\cdot)$ is on I, not the other way around.

3D reconstruction and novel view rendering. We first feed I_{outpaint} into SHARP [13] to obtain 3D Gaussian splats of the scene $\mathcal{G} = \mathcal{F}_{\text{recons}}(I_{\text{outpaint}})$. We then sample a random viewpoint i and render the corresponding Gaussian splatting (GS) image I_{gs} from the reconstructed 3D scene $\mathcal{G}$. This can be formulated as follows:

$$I_{\text{gs}} = \mathcal{F}_{\text{render}}\left(\mathcal{G}, R^\top, -R^\top \boldsymbol{t}\right), \tag{3}$$

where $R \in \mathbb{R}^{3\times3}, \boldsymbol{t} \in \mathbb{R}^3$ denote the rotation and translation from GS image I_{gs} to optimal image I[1]. Again, inverse rotation $R^\top$ and inverse translation $-R^\top \boldsymbol{t}$ are fed into $\mathcal{F}_{\text{render}}(\cdot)$ since the operation is on Gaussian splats from I_{outpaint}, not those from I_{gs}.

Image restoration. We use Qwen-Image-Edit [25] with a LoRA specialized for GS image restoration [4]. Specifically, the restoration model takes I_{gs} together with the outpainted image I_{outpaint} as input:

$$\tilde{I} = \mathcal{F}_{\text{restore}}\left(I_{\text{gs}}, I_{\text{outpaint}}\right), \tag{4}$$

where $\tilde{I}$ denotes the restored suboptimal image. This process removes holes and rendering artifacts in I_{gs} while preserving the viewpoint and scene content of the GS rendering.

Hierarchical filtering. As shown in Fig. 3, we adopt a three-stage hierarchical filtering scheme.

First, for outpainted image I_{outpaint}, we crop out the region from I_{outpaint} that corresponds to I and measure their similarity to verify consistency.

Second, we use an LLM [16] to filter GS images I_{gs} according to three criteria: subject consistency and completeness, ease of restoration, and depth consistency.

Third, for the suboptimal image $\tilde{I}$, we estimate the relative camera pose between I and $\tilde{I}$ using VGGT [18], and filter samples based on the rotation and translation-direction errors relative to the ground-truth rendering poses. Moreover, we use an LLM to further validate: (1) no visible

[1]The rotation matrix R in Eq. 3 is equivalent to the rotation quaternion $\boldsymbol{q}$ in Eq. 1.

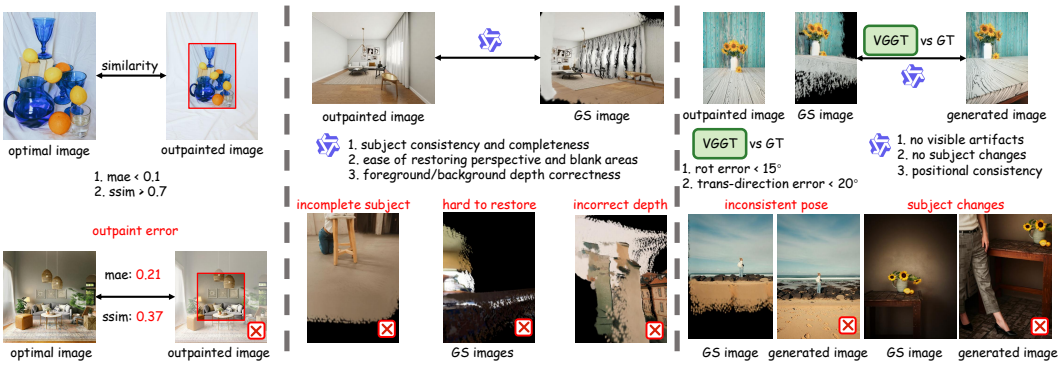


Figure 3: **Hierarchical filtering** for outpainting, novel views, and image restoration (left to right).

LLM Preference Rater Prompt

[Task]
You are a professional expert in photographic composition analysis. Your task is to compare two photos taken from different viewpoints (img1, img2) strictly based on photographic composition, and determine which photo has the better composition.

[Rule]
1. Identify the subject in img1 and img2 ...
2. Determine the genre and mode of expression of img1 and img2 ...
3. For each image, determine if the subject is clear, prominent, reasonably positioned, appropriately scaled, well coordinated with the background …
4. Explain the comparison process …

[Example]
{"input": img1, img2, "reasoning": "...", "preference": "..."}, …

Figure 4: Text prompt for LLM rater.

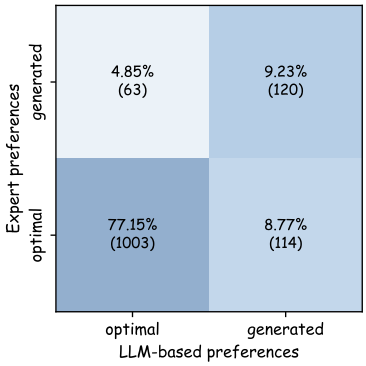


Figure 5: Confusion matrix between expert preferences and LLM preferences.

artifacts, distortions, or noise in $\tilde{I}$; (2) no subject changes compared with I_{outpaint}; and (3) pixel-level positional consistency with I_{gs}.

Discussion. Our pipeline is not restricted to a specific input format and can synthesize images from an arbitrary number of candidate viewpoints. This makes the dataset highly scalable: with sufficient computational resources, its scale can be continuously expanded. Moreover, as 3D reconstruction techniques continue to advance, or when multi-view reconstruction methods are adopted, the quality of the generated data is expected to further improve.

3.2 Benchmark

For performance evaluation of 3D viewpoint recommendation, we construct a dedicated benchmark using a subset of ViewRecDB-100K. Specifically, we ask experts to annotate 1,300 candidate images and select 1,000 for which I is confirmed to have better photography composition than the corresponding $\tilde{I}$.

LLM-based viewpoint rater. Since photography composition quality is difficult to quantify precisely, we formulate its evaluation as a pairwise comparison problem and introduce an LLM-based viewpoint rater with a CoT prompt to compare two images under professional photographic criteria. The prompt is refined based on professional photographer feedback and 100 annotated preference pairs with explanations. As shown in Fig. 4, the prompt guides the LLM through a step-by-step process in determining the viewpoint and photography composition preference.

When validated on 1,300 expert-annotated samples, the LLM judgment achieves an overall agreement of 86.38% with the expert annotations, indicating that the LLM-based viewpoint rater aligns well with human preferences.

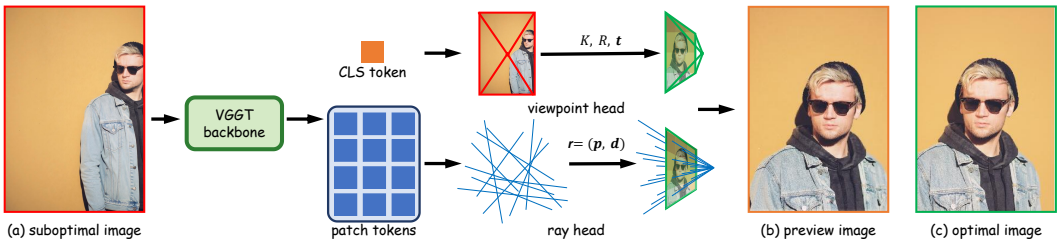


Figure 6: **ViewRecNet architecture.** Given a suboptimal image (a) as input, we use the VGGT backbone to generate the CLS token and patch tokens. The viewpoint head predicts the expert viewpoint based on the CLS token, while the ray head takes patch tokens as input and predicts camera rays for each patch. The predicted viewpoint can be used to generate the corresponding preview image (b), which is expected to mimic the optimal image (c).

4 The Viewpoint Recommendation Model

Given a suboptimal image $\tilde{I}$, the viewpoint recommendation model $f(\cdot)$ predicts:

$$(\hat{o}, \hat{\boldsymbol{v}}) = f(\tilde{I}), \tag{5}$$

where $\hat{o} \in \{0, 1\}$ is a binary orientation change indicator, and $\hat{\boldsymbol{v}} = (\hat{\boldsymbol{t}}, \hat{\boldsymbol{q}}, \hat{\boldsymbol{s}})$ represents the predicted viewpoint change from $\tilde{I}$ to the optimal image I, including $\hat{\boldsymbol{t}} \in \mathbb{R}^3$ and $\hat{\boldsymbol{q}} \in \mathbb{R}^4$ to denote the estimated camera pose change, and $\hat{\boldsymbol{s}} \in \mathbb{R}^2$ to estimate the FoV change.

Patch-as-ray representation. Since camera pose supervision for the recommended viewpoint $\boldsymbol{v}$ is sparse, we introduce dense geometric supervision following the cameras-as-rays paradigm [31]. Specifically, across all N_{patch} patches, each patch $i \in \{1, ..., N_{\text{patch}}\}$ corresponds to a ray:

$$\boldsymbol{r}_i = (\boldsymbol{p}_i, \boldsymbol{d}_i), \tag{6}$$

where $\boldsymbol{r}_i \in \mathbb{R}^6$ represents the camera ray passing through the center of the i-th patch, while $\boldsymbol{p}_i \in \mathbb{R}^3$ and $\boldsymbol{d}_i \in \mathbb{R}^3$ denote the origin and direction of the i-th ray, respectively.

The ray origin is set to the camera center for the optimal image I, which is given by $\boldsymbol{t}$. We denote the homogeneous center pixel coordinate of the image region corresponding to the i-th patch as $\boldsymbol{c}_i = (u_i, v_i, 1)^\top$. By back-projecting $\boldsymbol{c}_i$ into the world coordinate system, we obtain the ray direction:

$$\boldsymbol{d}_i = \frac{RK^{-1}\boldsymbol{c}_i}{\|RK^{-1}\boldsymbol{c}_i\|_2}, \tag{7}$$

where R denotes the rotation matrix from the suboptimal image $\tilde{I}$ to the optimal image I, and K denotes the camera intrinsic matrix. Since the camera intrinsics of the expert-suggested viewpoint are unavailable, we assume a fixed FoV for all input suboptimal images and use $\boldsymbol{s}$ to derive the FoV and intrinsic matrix K of the optimal viewpoint. In this way, patch-as-ray supervision provides dense constraints on both the camera intrinsic and extrinsic parameters.

4.1 Model Architecture

Our model consists of three main components: a VGGT backbone, a viewpoint head for viewpoint recommendation, and a ray head for ray-based dense supervision.

VGGT backbone. We adopt pretrained VGGT [18] as a feature backbone, and use its camera token (CLS token) and patch tokens for subsequent prediction.

Viewpoint head. Taking the CLS token as input, we use a two-layer MLP classifier to predict o, and adopt the same VGGT [18] camera head to predict $\boldsymbol{v}$.

Ray head. The camera ray map is predicted from the patch tokens using a three-layer MLP. Specifically, the MLP is shared across all patches and predicts $\boldsymbol{r_i}$ for the i-th patch token separately.

4.2 Training Objective

We train ViewRecNet end-to-end using a multi-task loss:

$$L = L_{\text{orientation}} + \alpha L_{\text{viewpoint}} + L_{\text{rays}}, \tag{8}$$

Table 2: Quantitative comparison. AUC@30 is the area under the accuracy-threshold curve up to 30°, computed based on RRA. LLM-Pref measures the pairwise preference between each baseline and our method. LLM-Imp measures whether the generated image is preferred over the input suboptimal image. Human-Pref reports the percentage of samples for which each method is selected as the best among our method and all baselines.

Method	Image-level Consistency				Pose-level Consistency		Preference		
	SSIM↑	LPIPS↓	PSNR↑	CLIP-I↑	AUC@30↑	Dist.↓	LLM-Pref.↑	LLM-Imp.↑[2]	Human-Pref.↑
CACNet	0.3616	0.7178	14.89	0.8608	0.3157	0.4843	34.40%	–	20.50%
Venus-Q	0.3584	0.7330	14.73	0.8623	0.2990	0.4475	25.60%	–	20.20%
PPC	0.3915	0.7167	15.10	0.6977	0.3060	0.5961	32.10%	53.10%	4.22%
SPAS	0.3751	0.7120	15.05	0.8421	0.4686	0.4587	35.70%	52.00%	10.60%
ViewRecNet	**0.3929**	**0.6919**	**15.31**	**0.8812**	**0.5271**	**0.3080**	reference	**67.80%**	**44.48%**

where $L_{\text{orientation}}$ adopts binary cross-entropy to supervise the landscape or portrait decision, $L_{\text{viewpoint}}$ adopts Huber loss on the 9-dimensional viewpoint descriptor $\boldsymbol{v}$ in Eq. 1 following [18], and L_{rays} is an $\mathcal{L}_2$ reconstruction loss over the 6-dimensional ray representation $\boldsymbol{r}_i$ (see Eq. 6) for each patch token i. We set $\alpha = 10$ by default. See Section B.1 for the sensitivity analysis.

4.3 Preview Image Rendering

To provide visual previews of the predicted viewpoint $\hat{\boldsymbol{v}}$, we use a pipeline similar to that described in Section 3.1. Specifically, given the suboptimal image $\tilde{I}$, we first reconstruct the 3D scene $\hat{\mathcal{G}}$ using $\mathcal{F}_{\text{recons}}(\cdot)$, then render the proposed view $\hat{I}_{\text{render}}$ (with the predicted $\hat{R}$ and $\hat{\boldsymbol{t}}$, not their inverse) using $\mathcal{F}_{\text{render}}(\cdot)$, and lastly restore the image $\hat{I}$ using $\mathcal{F}_{\text{restore}}(\cdot)$. No filtering is included.

5 Experiments

5.1 Experimental Setup

Training details. Selecting 100,000 samples randomly from ViewRecDB-100K as the training set, we train ViewRecNet end-to-end using two RTX 5090 GPUs for 5 epochs. AdamW is used with a maximum learning rate of 1×10^{-5}, weight decay of 0.05, and a linear learning-rate schedule with a warmup ratio of 0.1. The batch size is 16 per GPU.

Baselines. We compare ViewRecNet with two 2D cropping methods, CACNet [5] and Venus-Q [3], and two 3D viewpoint recommendation methods, PPC [28] and SPAS [9]. For PPC, we use the last frame of its generated view-adjustment video as the recommended view. Since the training data and code for SPAS are unavailable, we reimplement it following the published method and train it on our ViewRecDB-100K.

Evaluation metrics. We evaluate all methods on our benchmark from three aspects.

(1) Pose-level consistency. Since most models only output images without camera poses, we treat DA3 [10] as a reliable pose estimator. The pose recovered from the generated image is used as the predicted viewpoint, while the pose recovered from the optimal image is used as the ground-truth viewpoint. Following [19, 10, 9], we report AUC@30 based on RRA (relative rotation accuracy), along with Dist (average distance) between the two viewpoints.

(2) Image-level consistency. Following [28], we use CLIP-I, PSNR, SSIM, and LPIPS as image-level metrics. Since our model outputs a recommended viewpoint rather than an image, we modify the data generation pipeline to generate the corresponding preview image.

(3) Preference. We conduct LLM-based and human preference evaluations using LLM-Pref, LLM-Imp, and Human-Pref. LLM-Pref is the percentage of pairwise comparisons in which the LLM prefers the baseline result over the ViewRecNet result, while LLM-Imp measures whether the generated result is preferred over its suboptimal input. For Human-Pref, five raters independently evaluate all

[2]– denotes not applicable, as our LLM rater is designed for viewpoint-changing image pairs and may be unreliable for crop-only comparisons.

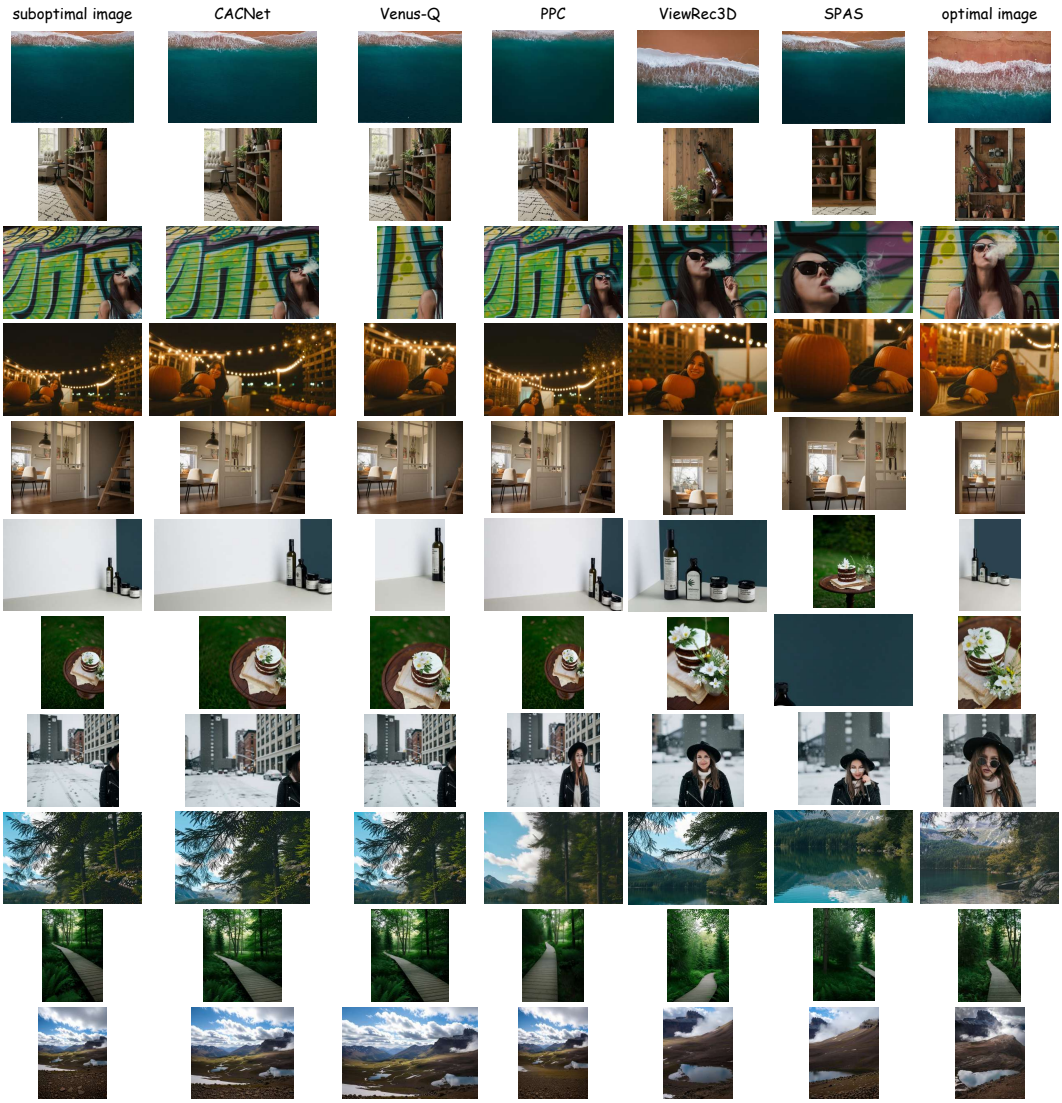


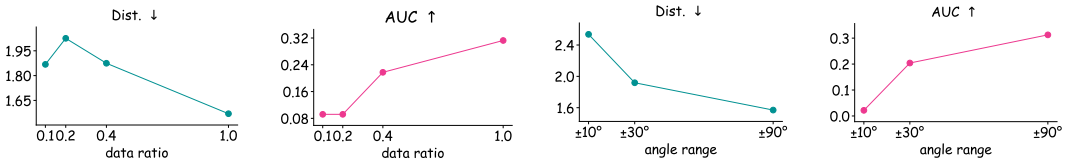
Figure 7: Qualitative comparison of 3D view recommendation results.

1,000 test samples by comparing the results of four baselines and ViewRecNet and selecting the one with the best composition.

5.2 Main Results

Quantitative analysis. As shown in Table 2, ViewRecNet outperforms all baseline methods across pose-level, image-level, and preference-based evaluations, indicating that it can learn the patterns of expert-selected viewpoints from our ViewRecDB-100K dataset and recommend viewpoints that better align with human compositional preferences. Compared with 2D cropping methods, ViewRecNet performs better in preference, further suggesting that 2D cropping is constrained by the original viewpoint and has limited ability to handle compositional adjustments that require substantial changes in spatial relationships. Compared with existing 3D viewpoint recommendation methods, particularly SPAS trained on the same ViewRecDB-100K dataset, ViewRecNet still achieves better performance, indicating that our method can learn more diverse viewpoint adjustments, thereby enabling more flexible viewpoint recommendation.

Qualitative analysis. As shown in Fig. 7, ViewRecNet generally yields views with more coherent and visually pleasing compositions than the baseline methods. It can correct tilted structural lines



(a) (b) (c) (d)

Figure 8: Impact of different training data generation settings. (a, b) show the impact of training data scale, and (c, d) show the impact of the camera angle change range.

Table 3: Ablation on hierarchical filtering.

	AUC@30 ↑	Dist. ↓
w/o filtering	0.2166	1.7421
w/ filtering	**0.3126**	**1.5697**

Table 4: Ablation on ray-based supervision.

	AUC@30 ↑	Dist. ↓
w/o ray	0.2654	1.7351
w/ ray	**0.3126**	**1.5697**

and improve visual flow, yielding more balanced and visually pleasing compositions. It can also coordinate the positions and relative proportions of different visual elements to form a more appropriate composition for the image content, for example by emphasizing the main subject or maintaining an overall balance among multiple elements. Compared with the two existing 3D baseline methods, ViewRecNet further demonstrates stronger viewpoint adjustment capability. Overall, ViewRecNet exhibits composition-aware capabilities in understanding scene structure and relationships among visual elements, representing a meaningful step toward AI photography.

5.3 Variant and Ablation Study

For all experiments in this section, we directly evaluate the viewpoints predicted by ViewRecNet against the ground-truth viewpoints using pose-level metrics. Specifically, AUC@30 is computed as the area under the joint accuracy curve for RRA&RTA (relative translation accuracy), over angular thresholds up to $30°$.

Impact of training data scale. We randomly select training subsets of 10k, 20k, and 40k samples and compare the resulting models with the model trained on the full 100k-sample training set. As shown in Fig. 8a and 8b, as the amount of training data increases, model performance generally improves (except for Dist fluctuation between 10k and 20k data samples). This verifies the effectiveness of our large-scale dataset.

Impact of rotation range in training data. For viewpoint changes, we limit the pitch/yaw ranges at three levels: $[-10, 10°]$, $[-30, 30°]$, and $[-90, 90°]$. Fig. 8c and 8d show that a broader range of camera pose changes in the training data helps the model learn a broader range of viewpoint adjustments, enabling it to better handle cases where the initial viewpoint differs substantially from the desired one. This further demonstrates the effectiveness of our data generation method.

Impact of hierarchical filtering. We evaluate the effect of hierarchical filtering on model performance in Tab. 3. Specifically, we combine the 97k samples rejected during the data generation process with ViewRecDB-100K to construct a mixed dataset, train ViewRecNet on this dataset, and compare it with the model trained only on ViewRecDB-100K. As shown in Tab. 3, after introducing these low-quality samples, model performance decreases noticeably despite the increase in training data scale. This suggests that, once the data scale is sufficiently large, data quality may be more important for viewpoint recommendation performance than simply increasing the amount of training data. Meanwhile, the results also show that our hierarchical filtering strategy is effective in identifying and removing low-quality samples.

Ray-based supervision. We construct a variant of ViewRecNet by removing the ray head and compare it with the full ViewRecNet to evaluate the effect of ray-based dense supervision. As shown in Tab. 4, removing the ray head leads to a performance drop, indicating that ray-based dense geometric supervision provides effective constraints during training and helps improve viewpoint recommendation performance.

5.4 Limitations and Broader Impact

ViewRecNet predicts only the recommended viewpoint, so image-level and preference-based evaluations rely on an additional preview generation pipeline involving 3D reconstruction, novel-view rendering, and image restoration. Errors introduced at these stages may affect the final preview and thus reduce evaluation accuracy. In addition, we use a single expert viewpoint as the optimal target, although photography composition may admit multiple valid solutions. Jointly predicting pose and image and modeling multiple plausible viewpoints are important directions for future work.

Regarding broader societal impact, this work has the potential to lower the barrier to high-quality photography by serving as an AI photography assistant that recommends more suitable 3D viewpoints to users. The proposed ViewRecDB-100K dataset provides large-scale paired data between expert-selected viewpoints and their corresponding suboptimal viewpoints, while ViewRecNet offers a complementary modeling framework, together providing new data and methodological foundations for research on AI photography.

6 Conclusion

In this paper, we presented ViewRecDB-100K, a viewpoint recommendation dataset constructed from expert-selected photos. To this end, we introduce a data generation method that renders high-quality image pairs from suboptimal viewpoints using 3D reconstruction. Moreover, a hierarchical filtering process is proposed for data quality control. On top of the dataset, we build a viewpoint recommendation model that uses ray-based dense geometric supervision. The proposed model performs strongly in both quantitative and qualitative evaluations and achieves the highest preference rate in the human evaluation study, supporting the effectiveness of the overall approach.

7 Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. U25A20434). We would like to thank our labmates for their valuable discussions and constructive feedback.

References

[1] 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 *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 226–234. IEEE, 2017. 2, 3

[2] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to compose with professional photographs on the web. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 37–45, 2017. 2

[3] Tianxiang Du, Hulingxiao He, and Yuxin Peng. Venus: Benchmarking and empowering multimodal large language models for aesthetic guidance and cropping. *arXiv preprint arXiv:2602.23980*, 2026. 3, 7

[4] dx8152. Qwen-Image-Edit-2511-Gaussian-Splash. https://huggingface.co/dx8152/Qwen-Image-Edit-2511-Gaussian-Splash. 2, 4

[5] 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. 3, 7

[6] James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning subject-aware cropping by outpainting professional photos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2175–2183, 2024. 2, 3

[7] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[8] 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. 3

[9] Jiawan Li, Fei Zhou, Zhipeng Zhong, Jiongzhi Lin, and Guoping Qiu. Towards smart point-and-shoot photography. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28242–28251, 2025. 2, 3, 7

[10] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 3, 7

[11] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or. Optimizing photo composition. In *Computer graphics forum*, volume 29, pages 469–478. Wiley Online Library, 2010. 2

[12] 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*, 2025. 3

[13] 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. *arXiv preprint arXiv:2512.10685*, 2025. 3, 4

[14] David Nordström, Johan Edstedt, Fredrik Kahl, and Georg Bökman. MuM: Multi-view masked image modeling for 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21736–21747, 2026. 14

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 14

[16] Qwen Team. Qwen3-VL-235B-A22B-Thinking. https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Thinking. 2, 4

[17] Unsplash. https://github.com/unsplash/datasets. 2, 3, 13

[18] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 3, 4, 6, 7, 14

[19] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 7

[20] Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. SeedEdit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025. 3

[21] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5261–5271. IEEE, 2025. 3

[22] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024. 3

[23] 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 conference on computer vision and pattern recognition*, pages 5437–5446, 2018. 2, 3

[24] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised pre-training for 3D vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 14

[25] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 3, 4

[26] 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13294–13304. IEEE, 2025. 3

[27] 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. 2, 3

[28] Lujian Yao, Siming Zheng, Xinbin Yuan, Zhuoxuan Cai, Pu Wu, Jinwei Chen, Bo Li, and Peng-Tao Jiang. Photography perspective composition: Towards aesthetic perspective recommendation. *arXiv preprint arXiv:2505.20655*, 2025. 2, 3, 7

[29] 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. 3

[30] 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. 2, 3

[31] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024. 2, 6

A Data Generation and Dataset Analysis

A.1 Data Generation and Filtering Details

Our data generation pipeline involves outpainting, 3D reconstruction, novel view rendering, and image restoration. Each stage may introduce artifacts or geometric errors. We apply hierarchical filtering to remove unreliable samples at intermediate stages and after restoration. Below, we report stage-wise sample counts and examine changes in scene distributions associated with filtering.

Stage-wise filtering statistics. We use the first 212,000 photos from approximately 300,000 featured photos in the Unsplash dataset [17] to construct ViewRecDB-100K. Figure A1 reports the number of images at each stage. During novel view rendering, we randomly sample 200 candidate viewpoints per image and retain at most 10 rendered results that satisfy the basic constraints. Novel view filtering then retains at most two viewpoints per source photo, followed by restoration and final filtering. The resulting dataset contains 104,704 valid pairs from 76,010 source photos, corresponding to a source-photo yield of 35.85%. These statistics show that filtering excludes a substantial number of intermediate samples from the final dataset.

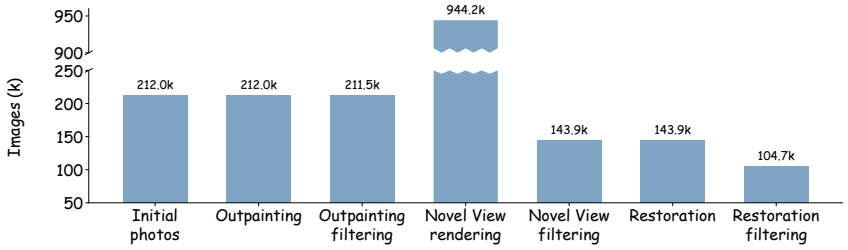


Figure A1: Stage-wise image counts in the data generation and filtering pipeline.

Scene distribution comparison. Figure A2 compares the scene distributions of all source photos and those contributing at least one valid pair after filtering. The proportions of nature and city scenes decrease slightly, while those of portraits, interiors, and still-life scenes increase slightly. All five categories still retain a sufficient number of samples, indicating that the hierarchical filtering process does not show an obvious bias toward removing or preserving any particular category. Therefore, the filtered dataset continues to cover diverse photographic scenarios while maintaining a reasonably balanced distribution across different scene types. These results further suggest that our hierarchical filtering strategy improves data quality without substantially reducing scene diversity, providing a reliable data foundation for subsequent model training and cross-scene generalization.

A.2 Viewpoint Distribution of ViewRecDB-100K

Figure A3 shows the yaw and pitch distributions in ViewRecDB-100K. Yaw changes span a broad range in both directions, whereas large pitch changes are relatively less frequent. Considering that substantial pitch adjustments are also uncommon in practical photography, this distribution is consistent with typical viewpoint changes in real-world shooting. In addition, since roll corresponds to an in-plane image rotation that can be easily handled by standard image processing operations, we do not explicitly model roll during data generation and set it to zero by default.

For translation, our representation is defined under a relative scale, so translation values across different scenes do not share a unified absolute physical scale. As a result, aggregating their global distribution is difficult to interpret in a physically meaningful way. We therefore focus primarily on the rotation distributions to characterize the range of viewpoint changes in the dataset, rather than further analyzing the global distribution of translation.

Overall, the angular distributions in ViewRecDB-100K cover the range of viewpoint adjustments commonly encountered in practical photography and retain sufficient variation across different directions. This indicates that the dataset provides diverse viewpoint changes and can reasonably simulate the typical viewpoint adjustment behavior of professional photographers.

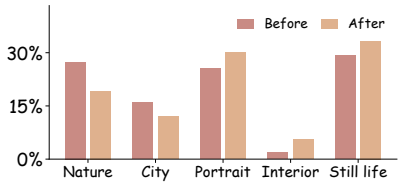


Figure A2: Scene distributions before and after hierarchical filtering.

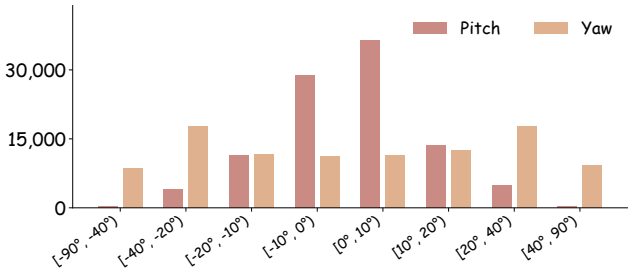


Figure A3: Rotation distributions on ViewRecDB-100K.

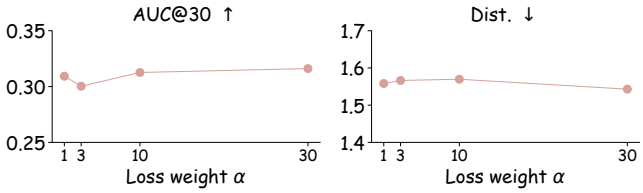


Figure B1: Sensitivity to the loss weight α.

Table B1: Comparison of different backbones.

Backbone	AUC@30↑	Dist.↓
DINOv2	0.2499	1.8417
CroCo	0.2092	2.0687
MuM	0.2524	1.9104
Frozen VGGT	0.0081	2.5081
VGGT (Ours)	**0.3126**	**1.5697**

B ViewRecNet Training and Model Analysis

B.1 Sensitivity to the Loss Weight

We investigate the sensitivity of ViewRecNet to the loss weight α, which balances the viewpoint loss against the orientation and ray reconstruction losses in the training objective. Figure B1 reports the results for $\alpha \in \{1, 3, 10, 30\}$.

Across the tested values, AUC@30 ranges from 0.3003 to 0.3161, while Dist ranges from 1.5430 to 1.5697. These results suggest that ViewRecNet is not highly sensitive to α within the evaluated range. Although $\alpha = 30$ achieves the best results on both metrics, its improvement over the default setting is modest: 0.0035 in AUC@30 and 0.0267 in Dist. We retain $\alpha = 10$ as the default setting for consistency across experiments.

B.2 Impact of Different ViewRecNet Backbones

We further investigate the impact of the pretrained backbone on viewpoint recommendation. Specifically, we replace VGGT [18] with DINOv2 [15], CroCo [24], and MuM [14] while keeping the prediction heads and training settings unchanged. All backbones are fully fine-tuned during training. We additionally evaluate a frozen VGGT variant, where the backbone parameters are kept fixed and only the prediction heads are trained.

As shown in Table B1, the frozen VGGT variant performs substantially worse than the fully fine-tuned model, indicating that the capability of ViewRecNet does not simply come from directly inheriting the pretrained representations of VGGT. Instead, task-specific adaptation of the backbone is important for learning viewpoint recommendation. Meanwhile, different pretrained backbones exhibit clearly different performance after being trained under the same framework. Among them, VGGT achieves the best results on both metrics, suggesting that its pretrained geometric representations provide a stronger initialization for this task. These results show that ViewRecNet benefits from both task-specific training and the choice of geometric foundation model.

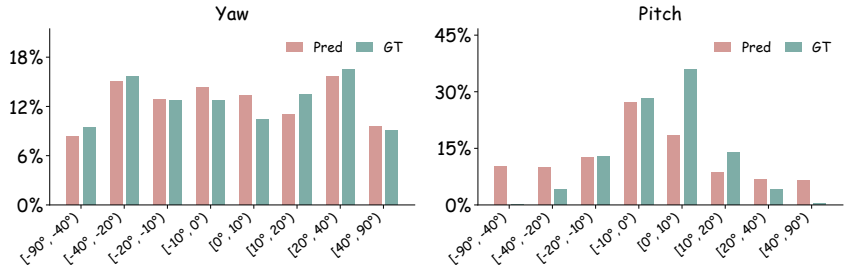


Figure C1: Distribution comparison of predicted (Pred) and ground-truth (GT) yaw and pitch.

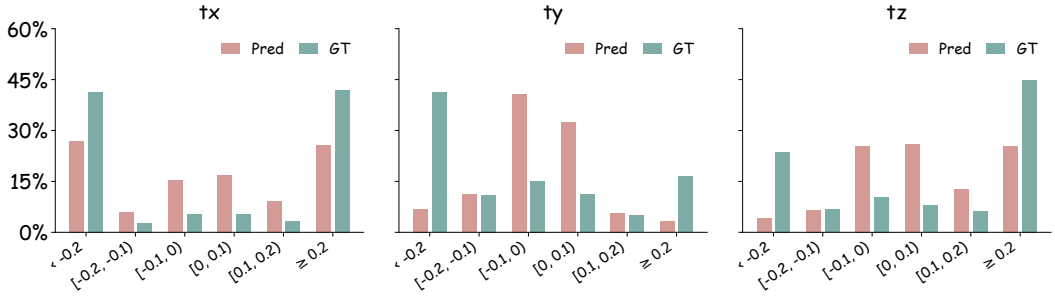


Figure C2: Distributions of predicted (Pred) and ground-truth (GT) translation components t_x, t_y, and t_z.

C Evaluation and Additional Results

C.1 Prediction Distribution Analysis

We further analyze the predicted and ground-truth viewpoints on our benchmark. Figures C1 and C2 compare the predicted and ground-truth yaw, pitch, and translation distributions. The predicted distributions retain substantial variation across different ranges and are not concentrated around zero or any fixed value, indicating that ViewRecNet does not collapse to a single fixed pose and instead produces diverse viewpoint adjustments according to the input.

C.2 Statistical Analysis of Human Preference

Table C1a reports Human-Pref with 95% confidence intervals, quantifying uncertainty in the estimated preference rates. Table C1b summarizes inter-rater agreement. The positive Fleiss' κ indicates agreement beyond chance, although participants do not consistently select the same result. This variation is consistent with the subjective nature of photographic composition, where multiple viewpoints may be visually appealing. Besides, the preference rates and agreement statistics provide complementary perspectives: the former summarize aggregate preferences across methods, while the latter characterize consistency among individual raters.

Table C1: Human preference analysis.

(a) Human preference with 95% CIs.

Method	Human-Pref. ↑	95% Confidence Interval
CACNet	20.50%	[18.76%, 22.28%]
Venus-Q	20.20%	[18.64%, 21.78%]
PPC	4.22%	[3.58%, 4.88%]
SPAS	10.60%	[8.98%, 12.26%]
ViewRecNet	**44.48%**	[42.42%, 46.60%]

(b) Inter-rater agreement.

Metric	Value
Observed agreement	0.5402
Expected agreement	0.2937
Fleiss' κ	0.3490

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: The abstract and Introduction clearly state the main claims. Our proposed contributions are derived from a rigorous analysis of the experimental results in Sec. 5 and are highly consistent with those results. In addition, we also clarify the assumptions and scope: we assume expert photographs to be optimal images, and the method is applicable to AI photography.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: Sec. 5.4 explicitly discusses the limitations of our method. Our model predicts camera poses rather than images from the predicted viewpoints, which means that image-level evaluation of the model is inevitably affected by errors introduced by the preview process in Sec. 4.3.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: This paper introduces a dataset, a benchmark and a model and reports experimental results; it does not present theoretical results such as theorems, propositions, or formal proofs.

   Guidelines:

   - The answer [N/A] 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: We describe the model architecture, training configuration, and evaluation protocol in the paper and supplementary material. We have publicly released our dataset, benchmark, and code to enable reproduction of the main experimental results.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: The code is released at the GitHub URL given in the abstract. The dataset is also publicly available online.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The training configuration and dataset split are described in detail in Sec. 5.1. The specific training data can be found in the dataset we release.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: We report 95% confidence intervals for the human preference rates in Table C1a in Appendix C.2, quantifying uncertainty in the preference estimates that support our main claims about compositional quality. We also report inter-rater agreement in Table C1b. These analyses concern human evaluation; we do not report uncertainty across training runs or random seeds.

Guidelines:

- The answer [N/A] 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: [Yes]

   Justification: Sec. 5.1 provides the type of GPU.

   Guidelines:

   - The answer [N/A] 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 this paper conforms to the NeurIPS Code of Ethics. The experiments involving human participants were conducted with appropriate consent and anonymization. No sensitive personal information is disclosed, and the study does not involve harmful or deceptive practices.

   Guidelines:

   - The answer [N/A] 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: [N/A]

Justification: The work focuses on a technical method for AI photography and automatic viewpoint selection. It is not directly tied to real-world deployment or decision-making systems that affect individuals or groups. Therefore, we do not identify direct societal impacts beyond its application value for AI-photographer systems.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release data or models that have a high risk for misuse, such as pre-trained language models, image generators, or large-scale unsafe scraped datasets. The released dataset and training code are intended for reproducing the reported AI photography experiments. We have obtained permission for the image data used and follow the corresponding terms of use, and we do not identify apparent dual-use or safety risks requiring additional safeguards.

Guidelines:

- The answer [N/A] 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 uses existing assets, including the Unsplash dataset and the SHARP, VGGT, and Qwen-Image-Edit models. We have obtained permission to use the Unsplash dataset and strictly follow its terms of use. We also properly credit the creators of SHARP, VGGT, and Qwen-Image-Edit, and use these models in accordance with their respective licenses and usage requirements.

Guidelines:

- The answer [N/A] 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 introduces new assets, including a dataset, a benchmark and code for reproducing the reported experiments. We provide documentation alongside the released assets, including dataset details, data splits, training configuration, usage instructions, licenses, and limitations. The released assets are anonymized at submission time where applicable.

Guidelines:

- The answer [N/A] 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 raters**

Question: For crowdsourcing experiments and research with human raters, 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: [No]

Justification: The paper involves human annotations or evaluations, but does not include the full text of instructions given to participants, screenshots of the annotation interface, or details about compensation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human raters.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human raters, 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 raters**

    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: [No]

    Justification: The paper involves research with human raters, but it does not describe IRB approval or an equivalent review process. The study does not collect sensitive personal information, but the approval status is not explicitly reported.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human raters.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human raters 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 rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes the use of LLMs as part of the experimental pipeline. Their roles, input-output behaviors, and impacts on the evaluation are specified in the method and experimental sections.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.