Aesthetic Camera Viewpoint Suggestion with 3D Aesthetic Field

Sheyang Tang[1], Armin Shafiee Sarvestani[1], Jialu Xu[1], Xiaoyu Xu[2], Zhou Wang[1]
[1]University of Waterloo, [2]City University of Hong Kong
{sheyang.tang,a5shafie,j565xu,x423xu,zhou.wang}@uwaterloo.ca

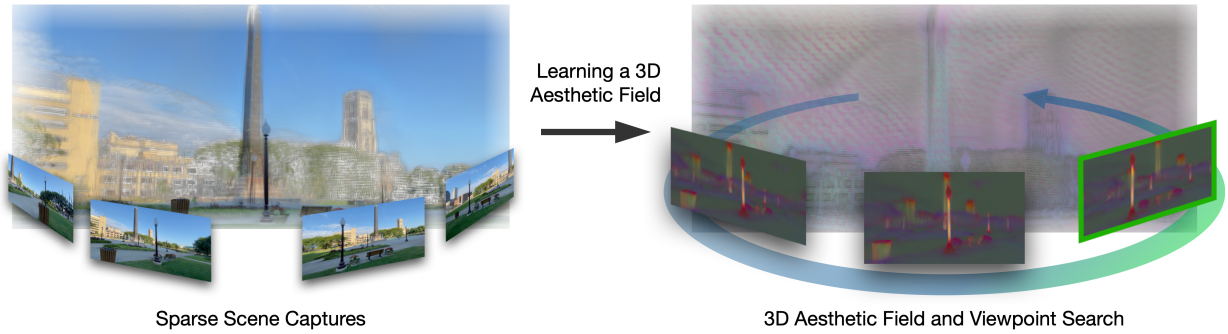


Figure 1. Given sparse scene captures (*left*), our method learns a 3D aesthetic field that encodes spatially varying aesthetic cues. This field enables geometry-grounded aesthetic reasoning in 3D, allowing efficient discovery of appealing camera viewpoints (*right*).

Abstract

The aesthetic quality of a scene depends strongly on camera viewpoint. Existing approaches for aesthetic viewpoint suggestion are either single-view adjustments, predicting limited camera adjustments from a single image without understanding scene geometry, or 3D exploration approaches, which rely on dense captures or prebuilt 3D environments coupled with costly reinforcement learning (RL) searches. In this work, we introduce the notion of 3D aesthetic field that enables geometry-grounded aesthetic reasoning in 3D with sparse captures, allowing efficient viewpoint suggestions in contrast to costly RL searches. We opt to learn this 3D aesthetic field using a feedforward 3D Gaussian Splatting network that distills high-level aesthetic knowledge from a pretrained 2D aesthetic model into 3D space, enabling aesthetic prediction for novel viewpoints from only sparse input views. Building on this field, we propose a two-stage search pipeline that combines coarse viewpoint sampling with gradient-based refinement, efficiently identifying aesthetically appealing viewpoints without dense captures or RL exploration. Extensive experiments show that our method consistently suggests viewpoints with superior framing and composition compared to existing approaches, establishing a new direction toward 3D-aware aesthetic modeling.

1. Introduction

Framing and composition play a central role in image aesthetics. The same 3D scene can appear engaging or awkward depending solely on the camera's viewpoint, as spatial relationships and perspective vary with viewing position. This makes aesthetics inherently *3D-dependent*. When framing a shot, experienced photographers intuitively reason about spatial layout by observing from multiple angles and forming an internal sense of scene aesthetics to anticipate how visual appeal changes with viewpoint [5]. This can be viewed as developing a mental map of aesthetic variation across viewpoints. Enabling machines to reason in a similar way—inferring spatially-varying scene aesthetics from a few observations to suggest appealing viewpoints—would benefit not only personal photography [11, 23] but also view selection/planning in VR/AR [28, 37] and autonomous systems [1, 32].

Existing approaches primarily focus on *single-view adjustments* [11, 14, 23] by predicting limited camera movements to refine a single image. Although they can enhance framing locally, they lack awareness of the underlying scene geometry, leaving their reasoning confined to a narrow neighborhood around the anchor view. Recent works attempt to incorporate broader viewpoint changes into aesthetic modeling [25, 35] using generative models. However, they rely on hallucinated content from a single view and thus cannot ensure geometric consistency with

real scenes. This highlights the need for a 3D-aware solution grounded in scene geometry for reliable aesthetic reasoning beyond observed views in 3D space.

In another line of work, *3D exploration* approaches explicitly operate in spatial environments, employing reinforcement-learning (RL) or genetic algorithms to search for appealing viewpoints [1, 22, 30, 32]. However, they assume access to dense, high-quality visual inputs from either real or readily available virtual 3D environments (*e.g.*, simulations, pretrained NeRFs [16]), which are costly to construct and require extensive data collection. Moreover, their RL-based solutions involve costly step-by-step exploration in the environment. These limitations call for a solution that alleviates the dependency on dense captures and enables efficient aesthetic reasoning without iterative physical adjustments in the real world.

To address these limitations, we propose to unify aesthetic perception with 3D geometry understanding by learning a *3D aesthetic field* that enables efficient inference of scene aesthetics from sparse observations. This field intrinsically encodes spatially varying aesthetic cues, supporting geometry-grounded aesthetic reasoning in 3D. Much like a skilled photographer who observes the scene from a few angles and develops a mental map of where the best angles lie, our system learns a 3D aesthetic field from sparse observations to reason about aesthetic variation across viewpoints and find appealing camera poses, as shown in Fig. 1. Specifically, we distill knowledge from a pretrained 2D aesthetic model [29] into a feedforward 3D Gaussian Splatting network [34], predicting per-Gaussian aesthetic features directly from sparse input views. Rendering aesthetic features from this field allows evaluation of aesthetic quality at novel viewpoints. The learned aesthetic field transforms viewpoint suggestion into a differentiable optimization problem, which we address by an efficient two-stage search pipeline: (1) we sample candidate viewpoints, and (2) locally refine them through gradient-based updates. By unifying aesthetic perception with geometric understanding, our method reasons beyond the observed views to efficiently identify optimal viewpoints within the scene, without the overhead of RL or dense captures for reconstruction. Extensive experiments demonstrate the superiority of our framework in aesthetic viewpoint suggestion. Our main contributions are:

- We introduce the task of *3D-aware* aesthetic viewpoint suggestion with *sparse observation*s, accounting for *3D-dependency* in aesthetic modeling without requiring *dense captures*.
- We propose a novel *3D aesthetic field* that unifies 2D aesthetic perception with 3D geometric understanding, modeling aesthetic variation across viewpoints.
- We develop an *efficient* two-stage search pipeline that combines coarse sampling with gradient-based refinement to efficiently discover appealing viewpoints.
- Extensive experiments on diverse datasets and input settings demonstrate the effectiveness of our framework for 3D-aware aesthetic viewpoint suggestion.

2. Related Works

2.1. Single-view Adjustment Methods

Early aesthetic optimization works focus on image cropping [6, 26, 29, 38] to reframe an existing photo to improve visual appeal. They predict cropping windows to maximize the aesthetic score but function only as post-processing once an image is captured. Later works extend aesthetic adjustment to virtual camera motion, reusing cropping datasets for learning camera adjustment that emulate crop movements. Su *et al*. [23] predict in-plane camera shifts for interactive composition refinement, while Li *et al*. [11] suggest camera rotations through an iterative feedback loop. Liu *et al*. [14] employ outpainting to suggest cropping windows beyond the image border, but their model assumes a fixed camera with an enlarged field of view, lacking true parallax changes. More recent efforts attempt to incorporate broader viewpoint variation. Uchida *et al*. [25] combine outpainting with single-image 3D reconstruction to enlarge the search space, and Yao *et al*. [35] synthesize poor-to-better perspectives via image-to-video generation. However, they rely on hallucinated content and cannot ensure geometric consistency with real scenes. Lacking awareness of the underlying 3D structure, single-view adjustment methods are restricted to local composition refinement around the original viewpoint. They struggle to discover more aesthetic angles that involve excluding or introducing scene elements, which requires 3D reasoning beyond observed views. In contrast, our method explicitly models spatially varying scene aesthetics, supporting geometry-aware aesthetic reasoning in 3D.

2.2. 3D-exploration Methods

3D-exploration methods reason directly in 3D environments. They employ RL or search algorithms to actively explore candidate viewpoints in real or simulated scenes. AutoPhoto [1] uses RL to navigate through 3D to find aesthetic views, while GAIT [32] trains an RL agent to generate indoor tours in simulation. ViewActive [30] uses geometric and semantic cues to optimize viewpoint selection for mesh objects. Skartados *et al*. [22] adopt a genetic algorithm to find aesthetic views in a pretrained NeRF scene. However, they rely on dense captures or prebuilt 3D assets that are costly to obtain, and their iterative RL procedures require iterative physical adjustments in real world. In contrast, our framework benefits from efficient inference of 3D scene aesthetics from sparse observations, and virtually reasons within the aesthetic field.

2.3. Feature Distillation in 3D Gaussian Splatting

Recent advances in 3D Gaussian Splatting [3, 9, 34] have enabled efficient, differentiable rendering and provided a versatile representation for encoding scene geometry and appearance. Building on this, several studies have explored distilling 2D semantic features [12, 18, 24, 40] into 3D Gaussian fields for segmentation tasks. These developments suggest the potential of transferring 2D knowledge into 3D representations. However, unlike semantic features that are largely viewpoint-invariant, aesthetic information is inherently viewpoint-dependent and remains largely unexplored in this context. Our work extends this paradigm by distilling aesthetic perception into 3D Gaussian fields to enable viewpoint-aware aesthetic modeling.

3. Proposed Method

Problem Statement. We aim to identify aesthetically pleasing viewpoints for a given 3D scene observed from sparse input views. We formulate this as finding the camera pose $\mathbf{P}^*$ that maximizes the predicted aesthetic score $score(\mathbf{P})$ over the continuous view space:

$$\mathbf{P}^* = \arg\max_{\mathbf{P}} score(\mathbf{P}). \tag{1}$$

Since direct optimization is intractable as it requires back-and-forth 2D projection and evaluation, we propose approximating this objective by learning a *3D aesthetic field*, which defines a continuous mapping from camera pose to aesthetic quality while grounded in scene geometry.

To learn this field, we *distill* aesthetic knowledge from a pretrained 2D aesthetic model into a feedforward Gaussian Splatting framework [3, 34] to reconstruct per-Gaussian geometry and aesthetic features from sparse input views. The distilled field enables rendering of aesthetic features at novel viewpoints, which are evaluated by an aesthetic decoder to assess framing and composition. To efficiently explore the 3D scene for viewpoint suggestion, we introduce a two-stage search pipeline: (1) coarse sampling of candidate viewpoints along and around the input trajectory followed by aesthetic scoring to select top candidates, and (2) gradient-based optimization for local refinement. The overall framework is illustrated in Fig. 2.

3.1. Distilling the 3D Aesthetic Field

Feedforward 3D Gaussian Splatting. We build our framework upon a feedforward Gaussian Spatting model [34] that predicts per-pixel Gaussian representation from sparse input views in a single forward pass. Specifically, given input views $\{\mathbf{I}^i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^N$ and corresponding camera poses $\{\mathbf{P}^i \in \mathbb{R}^{3\times 4}\}_{i=1}^N$, a multi-view transformer extracts multi-view features $\{\mathbf{F}^i_{mv}\}_{i=1}^N$. The features are then fused across neighboring views using plane-sweep [4, 33] aggregation and enriched with monocular depth cues to predict per-pixel depth, which determines the center $\boldsymbol{\mu}$ of each 3D Gaussian. A DPT head [19] finally regresses the remaining per-Gaussian parameters including covariance $\boldsymbol{\Sigma}$, opacity α, and color $\mathbf{c}$. Novel view images can be rendered from the 3D Gaussians via Gaussian Splatting [9]. This backbone provides a differentiable and efficient 3D representation, which we extend to construct the aesthetic field.

Direct RGB Scoring and Its Instability. Given the feedforward Gaussian Splatting backbone, a straightforward solution for viewpoint suggestion is to render novel views and directly score them using a pretrained aesthetic model. However, this naive approach faces two key challenges. *First*, aesthetic models are highly sensitive to small pixel variations, producing fluctuating scores even for adjacent views (see Fig. 3). This is because existing datasets lack annotations across nearby viewpoints, so the models haven't seen such variations during training. *Second*, rendering artifacts at novel views can bias aesthetic predictions and misguide optimization (see Fig. 3). To address them, we shift from pixel-level scoring to feature-level reasoning by distilling aesthetic representations into the 3D Gaussians. Operating in this latent feature space enhances robustness to low-level artifacts and enforces multi-view spatial consistency, leading to smoother aesthetic variations across nearby views. See Sec. 4.2 for further details.

Aesthetic Feature Distillation. We distill high-level aesthetic representations from a pretrained aesthetic teacher model [29]. Specifically, we select an intermediate layer as the distillation target and train our model to predict a per-Gaussian aesthetic embedding $\mathbf{f}_{aes}$, which can be decoded into aesthetic scores through the teacher's remaining layers after rasterization.

As shown in Fig. 2, on top of the backbone, we introduce lightweight aesthetic modules: a CNN aesthetic encoder, an aesthetic DPT head [19], and a transformer downsampler. The aesthetic encoder is from the teacher's feature extraction layers [29] and produces multi-scale aesthetic features $\{\mathbf{F}^i_{aes}\}_{i=1}^N$ from the input images. These features are fused with multi-view features $\{\mathbf{F}^i_{mv}\}_{i=1}^N$ from the backbone, and processed by the aesthetic DPT head [19] to regress per-Gaussian aesthetic embeddings $\mathbf{f}_{aes}$. Combined with other Gaussian attributes $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)$, these embeddings are rendered into aesthetic feature maps $\widehat{\mathbf{F}}_{pred}$ at novel views through the same rasterization pipeline as RGB rendering. To reduce storage and rasterization overhead, each $\mathbf{f}_{aes}$ is a compact 32-dim vector instead of 512-dim as the teacher feature. Since the teacher feature maps $\mathbf{F}_{gt}$ are much smaller and deeper than $\widehat{\mathbf{F}}_{pred}$, we employ a lightweight transformer downsampler to align them and obtain the final prediction $\mathbf{F}_{pred}$.

Furthermore, we condition the model on camera poses at both input and novel views, since aesthetic representations are inherently viewpoint-dependent. Incorporating pose in-

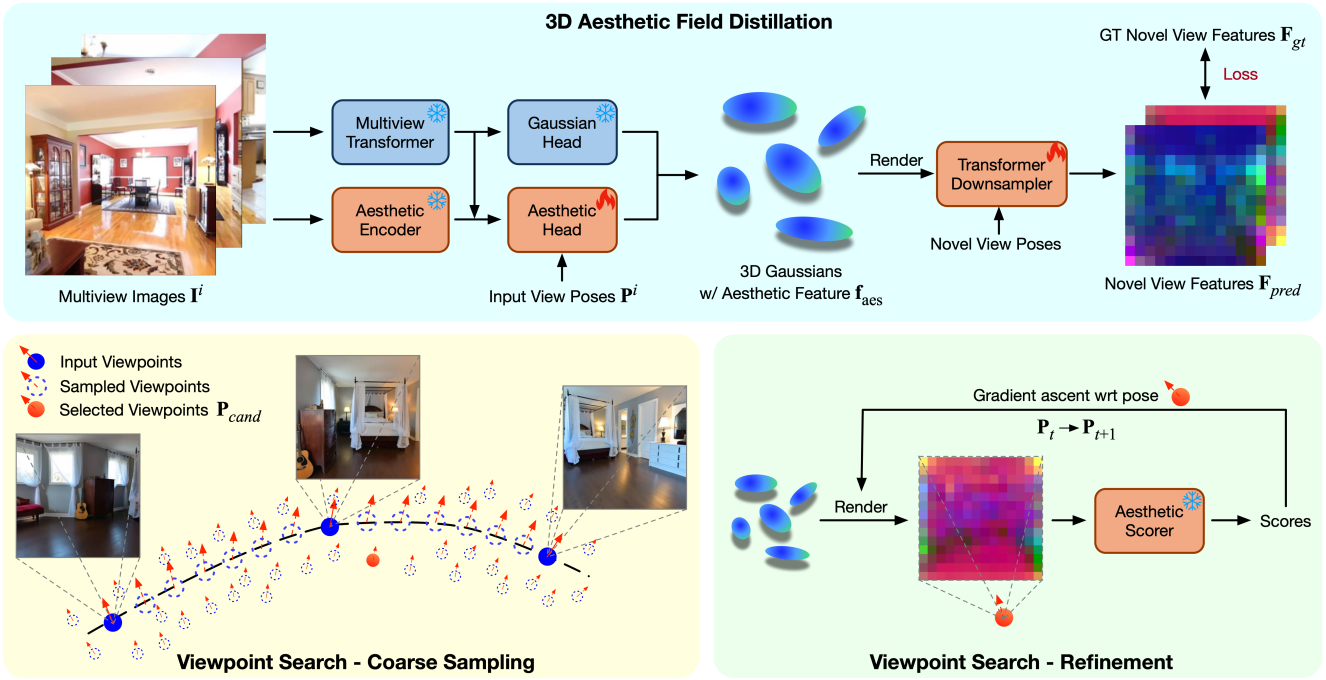


Figure 2. We distill aesthetic features into a feedforward Gaussian Splatting network (*top*). At inference, to search for aesthetic viewpoints, we adopt a two-stage pipeline: coarse sampling to find good candidates (*bottom left*) and local refinement by gradient ascent (*bottom right*).

formation allows the model to capture these dependencies.

Training Objective. The backbone multi-view transformer, DPT head, and aesthetic encoder are kept frozen to preserve consistent geometry predictions and 2D aesthetic perception, while the additional modules are trained end-to-end. We use mean-squared error (MSE) loss on the rendered feature maps $\mathbf{F}_{pred}$ at hold-out views, where ground-truth $\mathbf{F}_{gt}$ are obtained by feeding ground-truth images into the pretrained aesthetic model [29].

3.2. Viewpoint Search Pipeline

The learned 3D aesthetic field provides a continuous and differentiable mapping from camera pose to aesthetic score, enabling efficient search and optimization in viewpoint space. To efficiently explore this space, we adopt a two-stage coarse-to-fine search pipeline. In the first stage, candidate viewpoints are heuristically sampled around the input trajectories and scored to identify high-quality candidates. In the second stage, we perform gradient ascent on the selected candidates to locally refine their poses. The distilled aesthetic field yields a smooth score landscape (see Fig. 3 (a)), allowing stable gradient-based optimization (see Sec. 4.4). The final refined viewpoints are returned as aesthetic suggestions. The entire process is illustrated in the bottom of Fig. 2. We now describe each stage in detail.

Stage 1: Coarse Sampling. The goal of this stage is to efficiently identify promising viewpoint candidates for further refinement. We first connect the input views into a continuous camera trajectory by interpolating both camera positions and orientations, yielding a smooth path that covers the major observed viewpoints of the scene. Candidate viewpoints are then linearly sampled along this trajectory. For each sample, we further generate a local neighborhood of perturbed cameras with small in-plane shifts and directional jitters, allowing local exploration around the trajectory while maintaining scene focus. Each candidate viewpoint is rendered through the aesthetic field to obtain its aesthetic features, which are evaluated by the aesthetic decoder to produce a score. Finally, we select the top-K high-scoring candidates $\{\mathbf{P}^k_{cand}\}_{k=1}^K$ for the subsequent refinement stage. To enhance diversity, near-identical candidates are filtered out via a distance-based duplication check, ensuring that the selected set covers distinct viewpoints.

Stage 2: Gradient-based Refinement. Starting from the candidate viewpoints obtained in Stage 1, we perform local optimization directly on camera poses to further maximize their aesthetic scores. Specifically, each camera pose $\mathbf{P}^k_{cand}$ is parameterized by 3D translation and orientation, and updated through gradient ascent on the aesthetic score:

$$\mathbf{P}_{t+1} = \mathbf{P}_t + \eta \nabla_{\mathbf{P}} score(\mathbf{P}_t), \tag{2}$$

where η is the step size. In practice, we optimize a 5-dim vector comprising 3D translation and two rotation variables (yaw and pitch), as roll is rarely adjusted in typical scene captures. We use a fixed number of iterations for each can-

didate and retain the top-scoring refined viewpoints as the final suggestions.

4. Experiments

4.1. Experiment Setup

Datasets. We use RealEstate10k (RE10k) [41] and DL3DV [13] for training and evaluation. RE10k mainly contains indoor videos, while DL3DV involves more diverse scenes. Both have camera parameters at each frame. More details can be found in the Supplementary Material.

Implementation Details. We implement our framework in Pytorch [17], using DepthSplat [34] as the feedforward Gaussian Splatting backbone. We use the VEN [29] model as the teacher for distilling aesthetic representations. VEN is a CNN model, and we use features from the 23^{rd} layer($14 \times 14 \times 512$) as the distillation target. Following Xu *et al*. [34], we use image resolutions of 256×256 for RE10k and 256×448 for DL3DV. During training, we randomly sample 2 input views for RE10k and 2 to 6 input views for DL3DV, following the same protocol as Xu *et al*. [34]. Further implementation and training details are provided in the Supplementary Material.

Search Pipeline Configurations. At inference, we search for optimal views via a two-stage approach. In Stage 1, within each segment of the interpolated trajectory, we uniformly sample 16 camera poses, and around each we further sample 8 neighboring poses. We take top-2 candidates as input to the next stage. In Stage 2, we use the Adam [10] optimizer with step size 0.01 and update for 25 steps.

4.2. Aesthetic Prediction at Novel Views

We first evaluate our model's ability to predict aesthetic quality at unseen viewpoints to validate the learned 3D aesthetic field before applying it to viewpoint suggestion. Specifically, we compare our predicted aesthetic scores with ground-truth scores from the teacher aesthetic model [29], and benchmark against the RGB-scoring baseline. For each scene in the RE10k and DL3DV test splits, we use 2, 4, and 6 input views to build the aesthetic field and predict aesthetic scores for the remaining frames, using the entire video clips (about 250 frames each) for RE10k and 100 frames for DL3DV. Following Xu *et al*. [34], we train at 256×256 on RE10k and 256×448 on DL3DV, and evaluate at both resolutions to assess generalization.

Correlation to Teacher Scores. We compute the correlation between the predicted and teacher aesthetic scores at novel views using two standard metrics: the Pearson linear correlation coefficient (PLCC) and the Spearman rank-order correlation coefficient (SRCC), which are computed for each scene and then averaged. As shown in Tab. 1, across all settings, our aesthetic field achieves significantly higher correlation with the teacher than the RGB-scoring baseline, indicating that the distilled field produces more stable and faithful aesthetic predictions. As discussed ear-

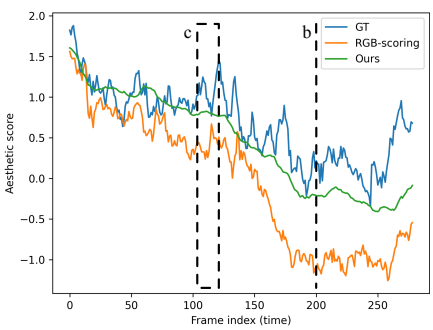
Methods	#Views	RE10k [41] PLCC	RE10k [41] SRCC	DL3DV [13] PLCC	DL3DV [13] SRCC
Baseline	2	0.657	0.628	0.326	0.307
Ours		**0.780**	**0.740**	**0.509**	**0.477**
Baseline	4	0.657	0.633	0.513	0.481
Ours		**0.796**	**0.758**	**0.722**	**0.682**
Baseline	6	0.745	0.701	0.580	0.553
Ours		**0.836**	**0.794**	**0.753**	**0.719**

(a) Image resolution 256×256.

Methods	#Views	RE10k [41] PLCC	RE10k [41] SRCC	DL3DV [13] PLCC	DL3DV [13] SRCC
Baseline	2	0.634	0.611	0.299	0.286
Ours		**0.733**	**0.704**	**0.479**	**0.445**
Baseline	4	0.596	0.578	0.552	0.535
Ours		**0.753**	**0.722**	**0.700**	**0.668**
Baseline	6	0.711	0.656	0.646	0.625
Ours		**0.821**	**0.764**	**0.737**	**0.707**

(b) Image resolution 256×448.

Table 1. Novel view aesthetic score correlation between predicted and ground-truth scores.



(a) Aesthetic score prediction. (b) GT & predicted frames.

(c) Nearby GT frames with fluctuation aesthetic scores.

Figure 3. (a) Aesthetic score predictions over consecutive frames. Our method produces scores closer to the ground truth while being smoother and more consistent across nearby views. The dashed line and box mark the regions visualized in (b) and (c), respectively. (b) Given the same aesthetic model and viewpoint, rendering artifacts in the predicted view (*bottom*) bias the RGB-scoring approach toward lower scores. (c) Ground truth scores fluctuate noticeably across nearly identical nearby views.

Methods	2 Input Views		4 Input Views		6 Input Views	
	VEN↑	SAMPNet↑	VEN↑	SAMPNet↑	VEN↑	SAMPNet↑
RE10k [41]						
Baseline	1.48	2.29	1.79	2.26	2.01	2.29
In-plane Shift*	1.52	2.31	1.81	2.36	2.10	2.41
Rotation*	1.78	2.38	1.95	2.42	2.13	2.45
UNIC [14]†	1.15	2.17	1.61	2.33	1.82	2.35
Uchida *et al*. [25]†	1.58	2.32	1.89	2.37	2.13	2.42
Ours	**1.89**	**2.40**	**2.03**	**2.45**	**2.20**	**2.49**
DL3DV [13]						
Baseline	2.08	2.23	2.31	2.26	2.47	2.30
In-plane Shift*	2.16	2.28	2.45	2.29	2.59	2.33
Rotation*	2.52	2.30	2.67	2.32	2.85	2.37
UNIC [14]†	2.07	2.21	2.35	2.27	2.48	2.30
Uchida *et al*. [25]†	2.34	2.25	2.68	2.32	2.81	2.36
Ours	**2.56**	**2.33**	**2.76**	**2.36**	**2.91**	**2.41**

Table 2. Aesthetic viewpoint suggestion evaluation on different datasets and with different numbers of input views. † indicates open-sourced methods adapted to our setting. * denotes approximations of non-open-sourced single-view methods in our setting. Note that such approximations may overestimate the actual results that those methods could deliver.

lier, direct RGB scoring is highly sensitive to rendering artifacts. Moreover, teacher scores themselves fluctuate across nearby ground-truth views with nearly identical compositions, so these quantitative results are best interpreted as *relative indicators* of stability and fidelity rather than absolute measures of accuracy. To better illustrate these effects and the advantages of our feature-distilled field, we present qualitative comparisons below.

Qualitative Comparisons. Figure 3 shows representative examples to illustrate the problems of RGB-scoring baseline and the superiority of our distilled aesthetic field. *First*, rendering artifacts in predicted novel-view images (*e.g*., noise, blurriness) can mislead the aesthetic model (Fig. 3 (b)). *Second*, ground-truth views with nearly identical content often yield fluctuating teacher scores (Fig. 3 (c)), revealing the inherent sensitivity of aesthetic models to pixel-level variations. As shown in Fig. 3 (a), these factors cause large oscillations in the RGB-scoring predictions (orange), whereas our aesthetic field (green) produces smoother and more consistent scores that closely follow the teacher trend (blue). Importantly, this stability is *not* achieved by explicit score smoothing, which would require arbitrary choices of window size or smoothness strength and lacks a principled basis without human ground-truth. Instead, our model *implicitly* enforces smoothness through feature distillation, where learning in the aesthetic feature space naturally regularizes the score landscape across neighboring viewpoints. These qualitative and quantitative results confirm that our distilled aesthetic field enables stable and consistent prediction of viewpoint-dependent aesthetics. More examples are provided in the Supplementary Material. The success in novel view aesthetic prediction forms the foundation for aesthetic viewpoint suggestion, which we evaluate next.

4.3. Aesthetic Viewpoint Suggestion

To evaluate our method on aesthetic viewpoint suggestion under sparse input views, we vary the number of input views over $[2, 4, 6]$ and assess the aesthetic quality of the suggested viewpoints.

Since no existing benchmarks account for viewpoint-dependent aesthetics, we adopt the test splits of RE10k and DL3DV to provide input views and use two aesthetic models (VEN [29] and SAMPNet [39]) to quantify framing and composition quality. Because the suggested views may not have corresponding ground-truth images, we reconstruct the scenes with dense inputs to obtain pseudo ground-truths for evaluation only, ensuring a fair and artifact-mitigated comparison without affecting the viewpoint suggestion process itself. As no prior work operates in this sparse-view setting, we compare against the baseline and single-view methods adapted to our evaluation protocol as detailed below.

Comparison Approaches. We first compare our method with the direct RGB-scoring baseline, which shares the same search pipeline as ours. To adapt single-view adjustment methods [11, 14, 21, 23, 25, 35] comparable in our setting, we treat each sparse input view as an anchor, around which we obtain a candidate based on their suggestions. This yields N suggestions for the N input anchors per scene, and we take the best as the final output. Among single-view methods, only UNIC [14] and Uchida *et al*. [25] are publicly available. For fair comparison, we adapt both by fixing their crop sizes as input resolution and map their outpainted suggestions to corresponding real viewpoints. See Supplementary Material for additional details. For

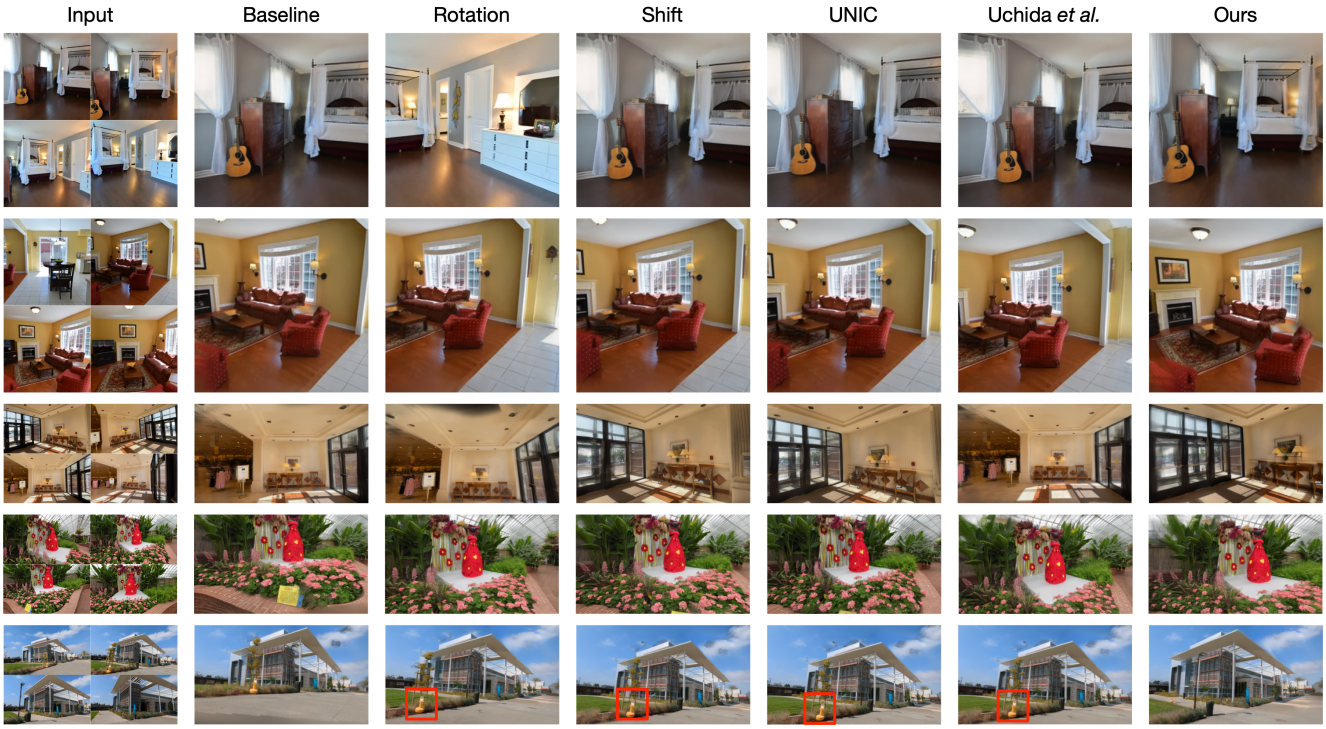


Figure 4. Aesthetic viewpoint suggestion examples with 4 input views in RE10k (*top 2 rows*) and DL3DV (*bottom 3 rows*). Red boxes in the last row show that single-view methods fail to remove distracting objects due to limited adjustment range. Zoom in for more details.

other methods that predict minor *in-plane shift* [21, 23] or *rotation* [11] around the anchor viewpoint, we approximate their behavior by exhaustively sampling and scoring nearby views with the aesthetic model. Yao *et al*. [35] cannot be compared as their video generation model is not available. 3D-exploration methods are not comparable as they require dense scene captures.

Quantitative Results. Table 2 reports the aesthetic viewpoint suggestion results. Our method consistently recommends views with higher aesthetic scores than all comparison approaches in different metrics. It also maintains superior performance across different input numbers, showing robustness to varying observation sparsity. Even with only two views, our model produces significantly better suggestions, showing its strong ability to reason about the underlying 3D scene aesthetics from minimal observations. As the number of input views increases, the suggested viewpoints achieve progressively higher scores, indicating that our model effectively exploits broader scene coverage to discover more aesthetically optimal perspectives. With more input views, the performance gap over single-view-based methods (*e.g*., *rotation*) narrows, likely because they can also explore sufficient viewpoint variation within their local neighborhoods. However, note that *inplane-shift* and *rotation* serve as an *upper bound* for approximating single-view methods in our setting, as we directly maximize target scores rather than relying on learned guidance.

Qualitative Comparisons. Figure 4 presents qualitative examples of the suggested viewpoints. Our framework identifies visually balanced and well-composed views that align with human aesthetic preferences. In contrast, single-view adjustment methods are restricted to small neighborhoods around the anchor views and often fail to capture optimal perspectives. For example, they may break the structural continuity of objects (*e.g*., bed frames in the first row) or struggle to remove distracting objects (*e.g*., marked by red boxes in the last row) from the view while our method can freely place the camera to enhance composition.

3D Visualization of Viewpoint Suggestions. Finally, we visualize the viewpoint sampling strategy in Fig. 5. Candidate viewpoints are sampled along and around the trajectory of input views, with colors indicating the predicted aesthetic scores. The results reveal how aesthetic quality varies across 3D space, and the corresponding renderings confirm strong alignment with human perceptual preferences.

4.4. Aesthetic Field Gradient Optimization Analysis

We analyze the learned aesthetic field through the lens of gradient-based optimization, and show that the distilled field enables stable and effective gradient ascent compared with the RGB-scoring baseline. We randomly select starting viewpoints and perform gradient ascent on both methods.

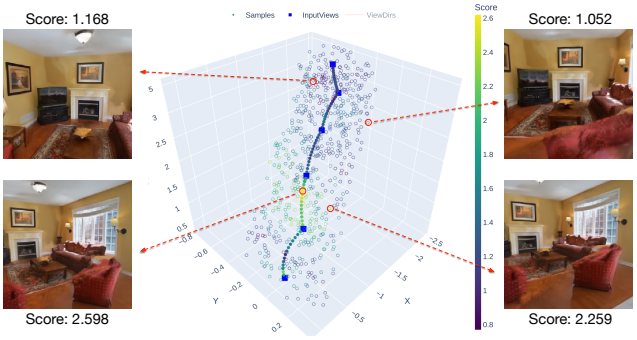


Figure 5. Visualization of sampled viewpoints colored by aesthetic score, with representative renderings shown alongside.

Methods	RE10k [41] ΔVEN↑	DL3DV [13] ΔVEN↑
Baseline	0.20	0.18
Ours	**0.46**	**0.43**

Table 3. Aesthetic score improvement with gradient ascent.

Figure 6. Gradient ascent results. Zoom in for a closer look.

For each scene in the test split, both approaches receive the same sparse input views, and the same random novel view is chosen as the initial camera pose for both. We run optimization for 25 steps and report the average score improvement ΔVEN in Tab. 3. Our method achieves consistent score increases in different datasets, confirming that the learned aesthetic field provides a well-behaved optimization landscape. In contrast, the RGB-scoring baseline suffers from unstable updates that often degrade the results. Figure 6 further illustrate this behavior. Our method converges toward balanced, aesthetically pleasing viewpoints, whereas RGB-based refinement fails to reliably update toward better views. See Supplementary Material for more examples.

4.5. Ablations

We conduct ablation studies to validate key architectural and configuration choices in our framework.

View-conditioning. We first examine the role of view-conditioning in aesthetic prediction for novel viewpoints. Tab. 4 shows that view-conditioning significantly improves aesthetic prediction accuracy at unseen views. This confirms that explicitly modeling viewpoint dependence is crucial for capturing aesthetic cues across diverse perspectives.

Methods	RE10k [41] PLCC	RE10k [41] SRCC	DL3DV [13] PLCC	DL3DV [13] SRCC
w/o. view-cond.	0.732	0.695	0.658	0.625
w. view-cond.	**0.796**	**0.758**	**0.700**	**0.668**

Table 4. Ablation of view-conditioning on novel view aesthetic prediction with 4 input views.

# Candidates	RE10k [41] VEN↑	DL3DV [13] VEN↑
1	1.96	2.57
2	2.03	2.76
3	2.05	2.78
# Steps	**ΔVEN↑**	**ΔVEN↑**
15	0.21	0.15
20	0.32	0.28
25	0.46	0.43
30	0.49	0.45

Table 5. Ablation of candidate numbers and refinement steps.

Search Configurations. We analyze the number K of candidates returned in coarse sampling and gradient ascent steps in refinement. As shown in Tab. 5, performance gains saturate around $K = 2$ and 25 refinement steps, which are therefore adopted as default in all experiments. Additional ablations are provided in the Supplementary Material.

5. Discussion & Conclusion

We present a novel framework for 3D-aware aesthetic viewpoint suggestion that learns a 3D aesthetic field via feedforward 3D Gaussian Splatting, enabling geometry-grounded aesthetic reasoning and efficient aesthetic viewpoint discovery from sparse input views. Despite its effectiveness, our framework still has room for improvement. First, it relies on camera poses to build the aesthetic field. While such information can be obtained from COLMAP [20] and is often available in robots/drones and smartphones [7], a pose-free variant would broaden the applicability of our method. This could be achieved by incorporating recent advances in pose-free methods [8, 36]. Second, the quality of the aesthetic field depends on the accuracy of the reconstructed geometry, which is affected by both the backbone's ability and the input coverage. While the former can be improved by stronger geometry backbones [27], the latter can be addressed by selecting views to ensure sufficient scene coverage [2, 31]. Third, our viewpoint search is constrained to regions supported by the initial observations. A promising extension is an active-perception loop that acquires additional views in promising directions to expand the aesthetic field and enlarge the feasible search space.

Aesthetic Camera Viewpoint Suggestion with 3D Aesthetic Field

Supplementary Material

S1. Datasets & Implementation Details

We conduct experiments on subsets of the RE10k [41] and DL3DV [13] datasets. For RE10k, we use approximately $6,000$ scene clips as the training split and 749 clips as the test split. For DL3DV, we adopt its 1K-5K subsets for training, which contain roughly $5,000$ scenes, and test on the 140 benchmark scenes.

During training, we sample 2 input views to reconstruct the aesthetic features of 12 target views in RE10k, and sample 2 to 6 input views to predict aesthetic features for 24 target views in DL3DV. We train all models for 50,000 steps with a batch size of 8. We use AdamW [15] with a learning rate of 0.0005 and a cosine annealing scheduler.

S2. Aesthetic Prediction at Novel Views

We provide additional qualitative examples of novel-view aesthetic prediction to further illustrate the two limitations of the RGB-based baseline approach discussed in the main paper. As shown in Fig. S1, the aesthetic model is biased by the rendering artifact, and exhibits high sensitivity to minor pixel-level variations across nearby views that are nearly identical. In contrast, our distilled aesthetic field faithfully follows the teacher's underlying trend, producing smoother and more consistent predictions across nearby frames.

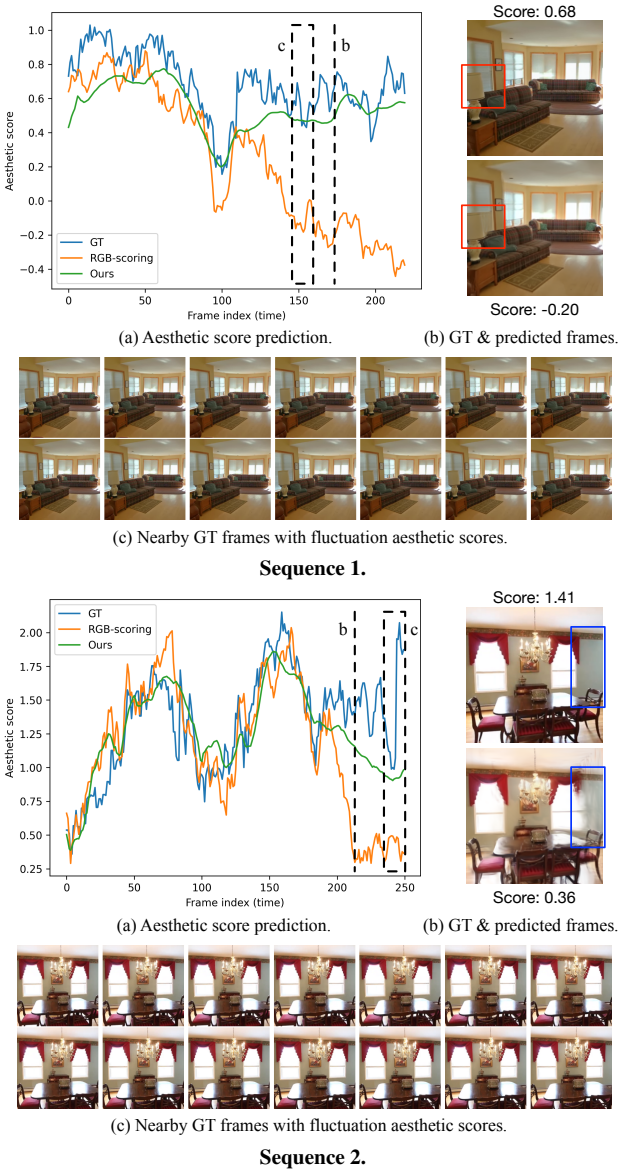
S3. Aesthetic Viewpoint Suggestion

We first describe how existing single-view baselines are adapted to our evaluation setting, followed by additional qualitative comparisons of the suggested viewpoints.

UNIC [33] predicts crops with varying sizes in an extrapolated image plane rather than real camera motions. For fair comparison, we fix its crop size as input resolution (256×256 in RE10k and 256×448 in DL3DV) and map the crop-center shifts to in-plane camera translations, thereby obtaining corresponding real viewpoints where we measure aesthetic qualities. Uchida *et al*. [25] first outpaint an image, reconstruct 3D point clouds from it, and then optimize camera poses and image aspect ratios in this outpainted 3D scene. We therefore also fix their output resolution and transform their suggestions to equivalent real viewpoints for evaluation.

As shown in Fig. S2, our method consistently suggests viewpoints with superior framing and composition than comparison methods.

Finally, additional 3D visualizations of the viewpoint sampling results are shown in Fig. S3. The results demonstrate the variation of aesthetic quality across 3D space, and the corresponding renderings confirm strong alignment be-



(a) Aesthetic score prediction. (b) GT & predicted frames.

(c) Nearby GT frames with fluctuation aesthetic scores.

Sequence 1.

(a) Aesthetic score prediction. (b) GT & predicted frames.

(c) Nearby GT frames with fluctuation aesthetic scores.

Sequence 2.

Figure S1. Additional novel view aesthetic prediction results. *In both examples:* (a) Aesthetic score predictions over consecutive frames. The dashed line and box mark the regions visualized in (b) and (c), respectively. (b) Given the same aesthetic model and viewpoint, rendering artifacts in the predicted view (*e.g.*, wrong color prediction as marked in red boxes in *top*, noisy predictions marked in blue boxes in *bottom*) bias the RGB-scoring approach. (c) Ground-truth scores fluctuate noticeably across nearly identical nearby views.

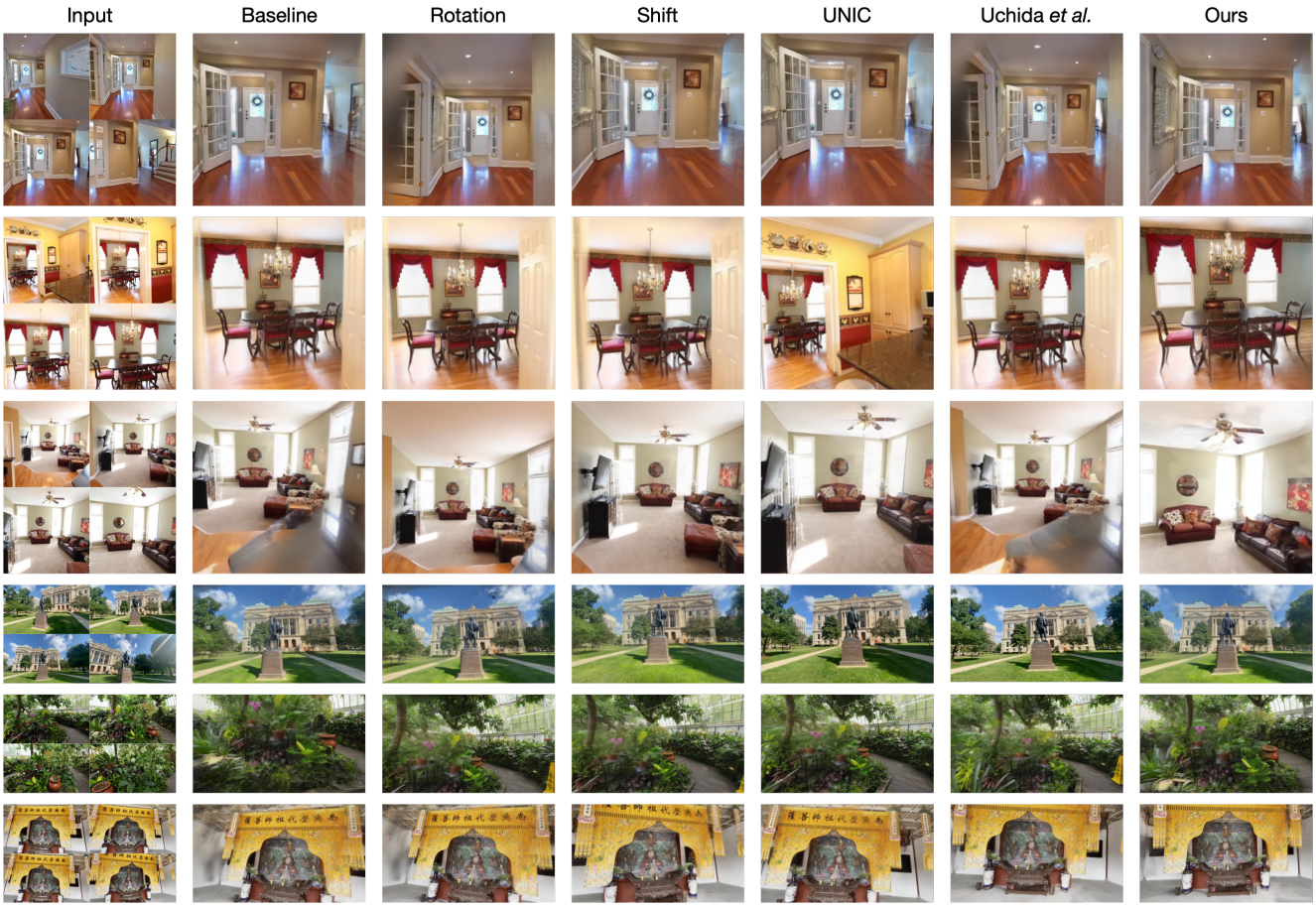


Figure S2. Additional results of aesthetic viewpoint suggestion.

tween our model's evaluations and human perceptual preferences.

S4. Aesthetic Field Gradient Optimization

We present additional results comparing gradient ascent optimization using the baseline approach and our method. As shown in Fig. S4, our method consistently converges toward more aesthetically pleasing viewpoints, whereas the baseline approach often fails to make meaningful progress and can sometimes degenerate.

S5. Ablations

We provide additional ablation experiments of the Stage 1 search configuration, focusing on two parameters: the number of samples along each input-view segment (S) and the number of neighbors further drawn around each segment sample (N). The total number of candidate viewpoints is approximately proportional to $S \times N$. We measure the aesthetic qualities of final suggestions via VEN [29]. As shown in Tab. S1, our method remains stable across a wide range of sampling densities, while generally achieving better performance with more samples. We adopt $S = 16$ and $N = 8$ as our default configuration, as this setting offers a good balance between suggestion quality and total number of samples being evaluated.

	$S = 4$	$S = 8$	$S = 16$	$S = 32$
$N = 4$	1.98	2.00	1.98	1.98
$N = 8$	1.98	1.94	2.03	2.02
$N = 16$	2.01	2.00	2.02	2.04
$N = 32$	2.02	2.03	2.04	2.06

Table S1. Ablation of the sampling configuration of search stage 1 on novel view aesthetic prediction. S denotes the number of samples along each segment of the interpolated input-view trajectory, and N denotes the number of neighboring viewpoints sampled around each segment-sample. All experiments use 4 input views on RE10k [41], and the aesthetic quality is measured by VEN [29].

References

[1] Hadi AlZayer, Hubert Lin, and Kavita Bala. Autophoto: Aesthetic photo capture using reinforcement learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 944–951. IEEE, 2021. 1, 2

[2] Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, and Jiangmiao Pang. Gennbv: Generalizable next-best-view policy for active 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16436–16445, 2024. 8

[3] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 3

[4] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE computer society conference on computer vision and pattern recognition*, pages 358–363. Ieee, 1996. 3

[5] Michael Freeman. *The Photographer's Eye Digitally Remastered 10th Anniversary Edition: Composition and Design for Better Digital Photos*. Routledge, 2017. 1

[6] 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. 2

[7] Apple Inc. Apple arkit. https://developer.apple.com/augmented-reality/arkit/, 2025. Accessed: 2025-11-13. 8

[8] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025. 8

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

[10] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[11] 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. 1, 2, 6, 7

[12] Qijing Li, Jingxiang Sun, Liang An, Zhaoqi Su, Hongwen Zhang, and Yebin Liu. Semanticsplat: Feed-forward 3d scene understanding with language-aware gaussian fields. *arXiv preprint arXiv:2506.09565*, 2025. 3

[13] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 5, 6, 8, 1

[14] Xiaoyu Liu, Ming Liu, Junyi Li, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Beyond image borders: Learning feature extrapolation for unbounded image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13023–13032, 2023. 1, 2, 6

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[18] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 3

[19] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3

[20] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 8

[21] Nan Sheng, Yongzhen Ke, Shuai Yang, Yong Yang, and Liming Chen. View adjustment: helping users improve photographic composition. *Multimedia Systems*, 30(5):293, 2024. 6, 7

[22] Evangelos Skartados, Mehmet Kerim Yucel, Bruno Manganelli, Anastasios Drosou, and Albert Saà-Garriga. Finding waldo: Towards efficient exploration of nerf scene spaces. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pages 155–165, 2024. 2

[23] Yu-Chuan Su, Raviteja Vemulapalli, Ben Weiss, Chun-Te Chu, Philip Andrew Mansfield, Lior Shapira, and Colvin Pitts. Camera view adjustment prediction for improving image composition. *arXiv preprint arXiv:2104.07608*, 2021. 1, 2, 6, 7

[24] Qijian Tian, Xin Tan, Jingyu Gong, Yuan Xie, and Lizhuang Ma. Uniforward: Unified 3d scene and semantic field reconstruction via feed-forward gaussian splatting from only sparse-view images. *arXiv preprint arXiv:2506.09378*, 2025. 3

[25] Taichi Uchida, Yoshihiro Kanamori, and Yuki Endo. 3d view optimization for improving image aesthetics. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 1, 2, 6

[26] Chao Wang, Li Niu, Bo Zhang, and Liqing Zhang. Image cropping with spatial-aware feature and rank consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10052–10061, 2023. 2

[27] 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. 8

[28] Zhan Wang, Qian Zhu, David Yip, Fugee Tsung, and Wei Zeng. Cinefolio: Cinematography-guided camera planning for immersive narrative visualization. *Visual Informatics*, page 100259, 2025. 1

[29] 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, 4, 5, 6

[30] Jiayi Wu, Xiaomin Lin, Botao He, Cornelia Fermuller, and Yiannis Aloimonos. Viewactive: Active viewpoint optimization from a single image. *arXiv preprint arXiv:2409.09997*, 2024. 2

[31] Wenhui Xiao, Rodrigo Santa Cruz, David Ahmedt-Aristizabal, Olivier Salvado, Clinton Fookes, and Leo Lebrat. Nerf director: Revisiting view selection in neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20742–20751, 2024. 8

[32] Desai Xie, Ping Hu, Xin Sun, Soren Pirk, Jianming Zhang, Radomír Mech, and Arie E Kaufman. Gait: Generating aesthetic indoor tours with deep reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7409–7419, 2023. 1, 2

[33] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13941–13958, 2023. 3, 1

[34] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16453–16463, 2025. 2, 3, 5

[35] 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. 1, 2, 6, 7

[36] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024. 8

[37] Faisal Zaman, Craig Anslow, and Taehyun James Rhee. Vicarious: Context-aware viewpoints selection for mixed reality collaboration. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*, pages 1–11, 2023. 1

[38] 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. 2

[39] Bo Zhang, Li Niu, and Liqing Zhang. Image composition assessment with saliency-augmented multi-pattern pooling. *arXiv preprint arXiv:2104.03133*, 2021. 6

[40] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 3

[41] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 5, 6, 8, 1, 2

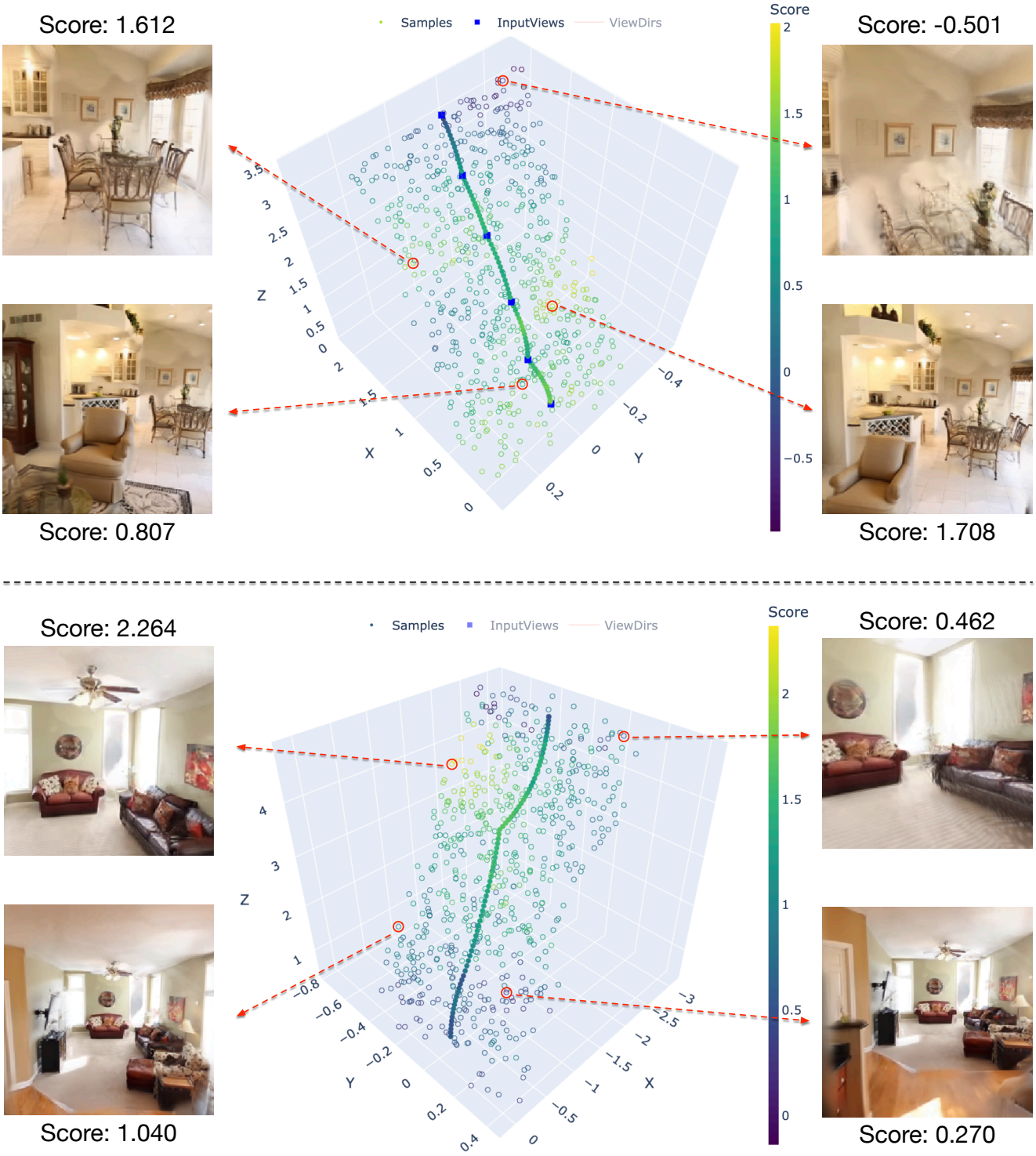


Figure S3. Additional visualizations of sampled viewpoints colored by aesthetic score, with representative renderings shown alongside.

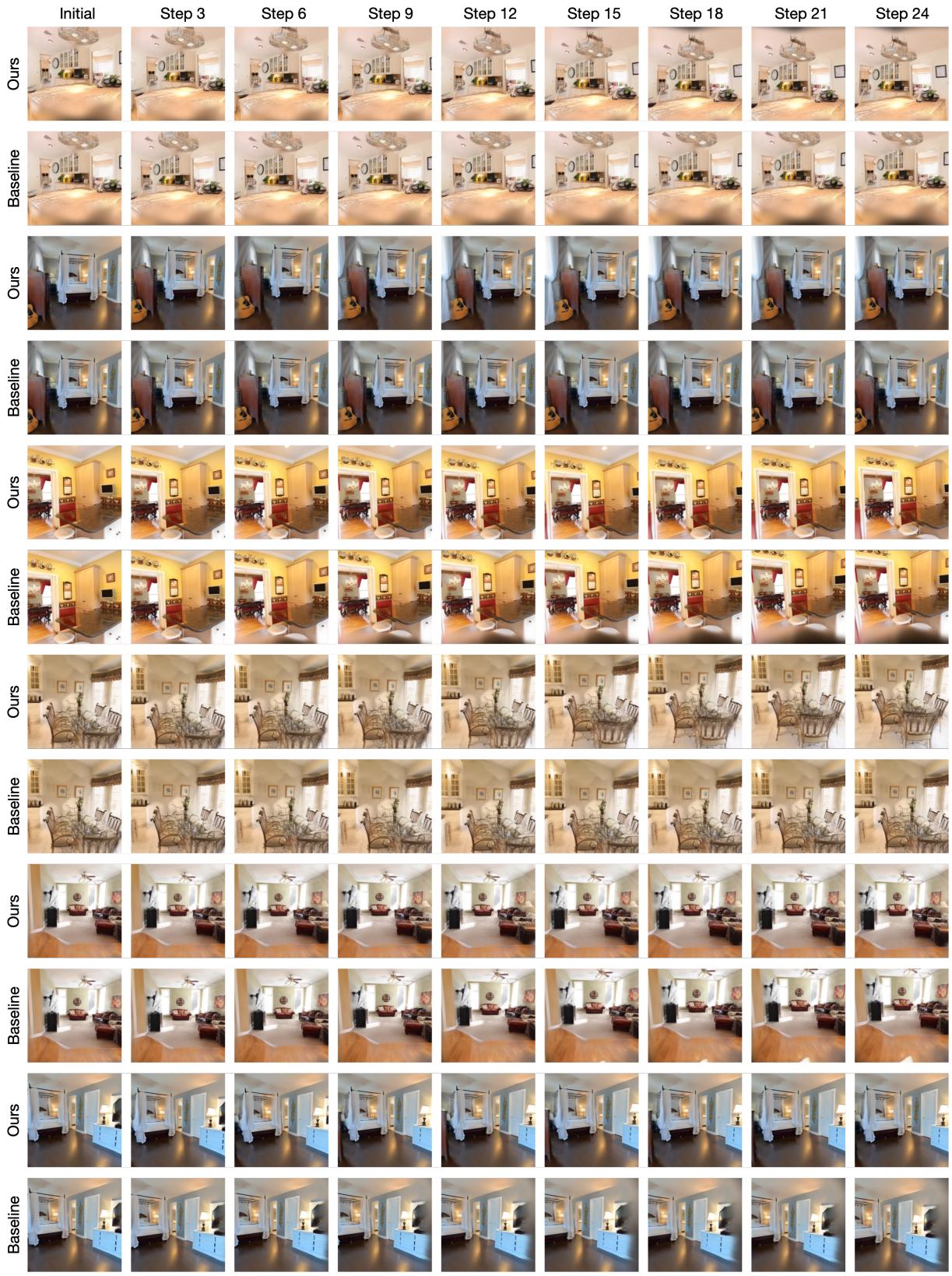


Figure S4. Additional gradient ascent results. Zoom in for a closer look.