# Aesthetic Camera Viewpoint Suggestion with 3D Aesthetic Field

## Supplementary Material

## S1. Datasets & Implementation Details

We conduct experiments on subsets of the RE10k [6] and DL3DV [1] 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 [2] 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.

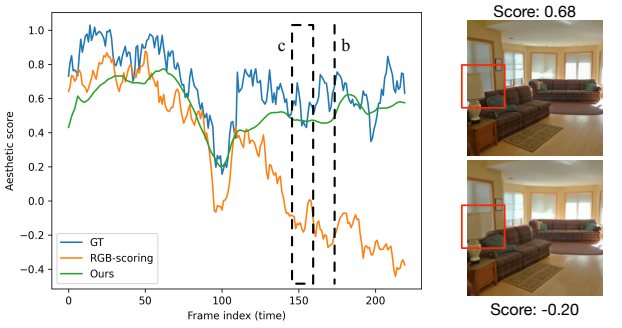
## S3. Aesthetic Viewpoint Suggestion

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

UNIC [5] 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 ($256\times256$ in RE10k and $256\times448$ 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*. [3] 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.

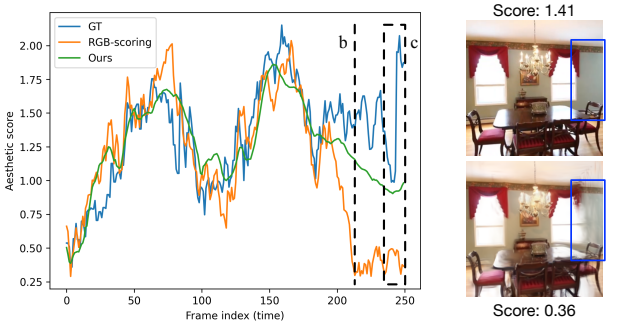
Finally, additional 3D visualizations of the viewpoint sampling results are shown in Fig. S3. The results demonstrate the variation of aesthetic quality across 3D space, and the corresponding renderings confirm strong alignment be-



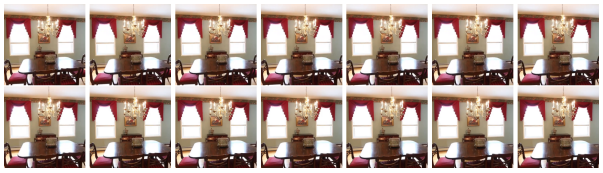

(a) Aesthetic score prediction. (b) GT & predicted frames.

(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.

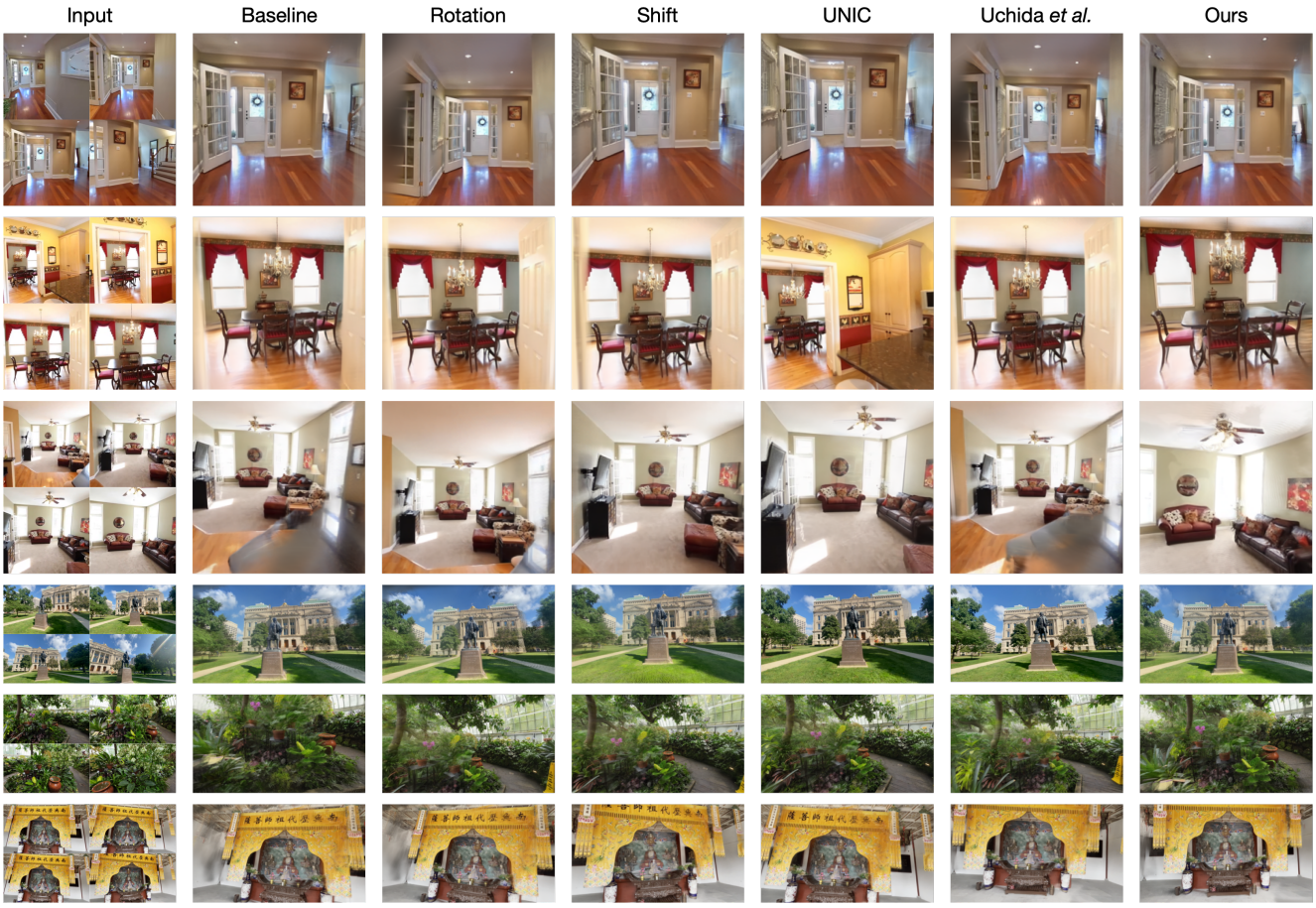



Figure S2. Additional results of aesthetic viewpoint suggestion.

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 [4]. 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 [6], and the aesthetic quality is measured by VEN [4].

## References


[1] 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. 1

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

[3] 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

[4] 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

[5] 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. 1

[6] 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. 1, 2

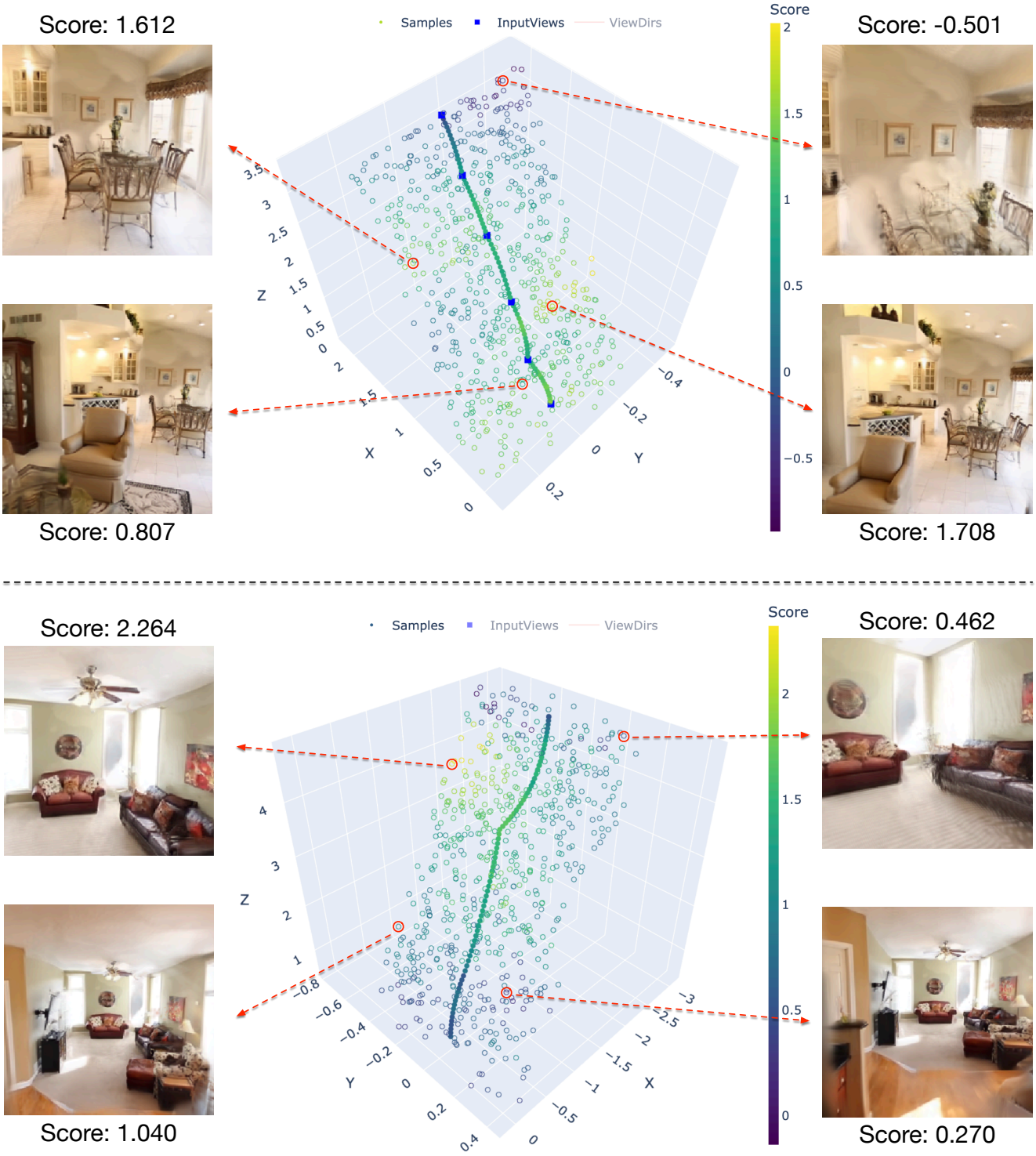



Figure S3. Additional visualizations of sampled viewpoints colored by aesthetic score, with representative renderings shown alongside.

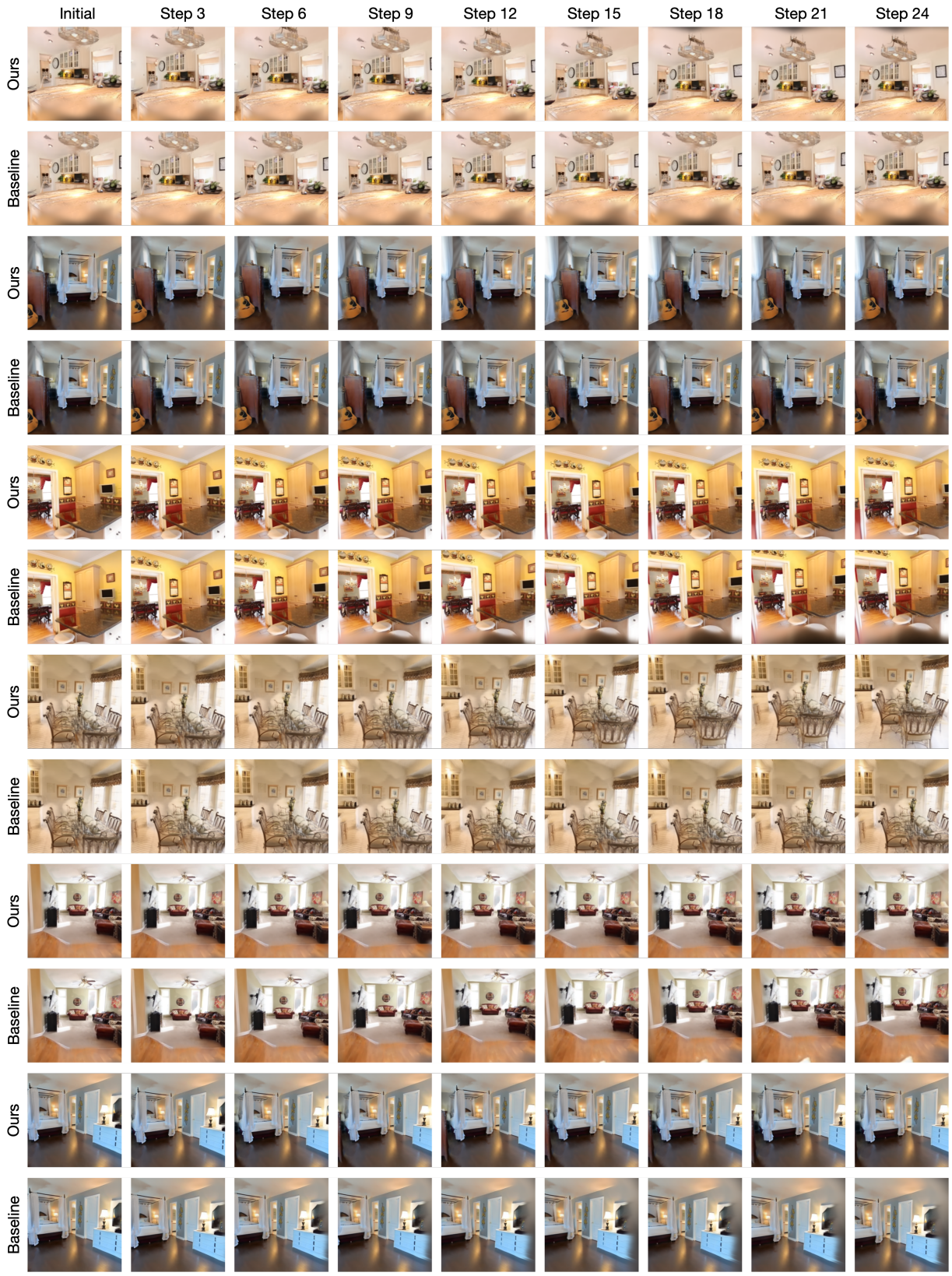



Figure S4. Additional gradient ascent results. Zoom in for a closer look.