Correspondence-Attention Alignment for Multi-View Diffusion Models

Supplementary Material

Appendix

This appendix presents additional experimental results and further details of our proposed method, CAMEO.

- Sec. A reviews the fundamentals of diffusion models.
- Sec. B describes the architecture of multi-view diffusion models in detail.
- Sec. C provides a detailed analysis of correspondence in multi-view diffusion models, including the analysis setup.
- Sec. D covers implementation details, including correspondence map derivation, and implementation of baseline models.
- Sec. E discusses how the model behavior and correspondence patterns evolve after training.
- Sec. F shows additional ablation studies.
- Sec. G presents additional qualitative results.
- Sec. H provides the results and implementation details of 3D reconstruction.
- Sec. I describes the limitations of CAMEO.
- Sec. J discusses future directions.

A. Preliminaries for diffusion models

Diffusion models [14, 42] are a class of generative models that learn data distributions by reversing a gradual noising process. Starting from clean data samples $x_0 \sim p_{\text{data}}(x)$, a forward process incrementally corrupts them with Gaussian noise to produce a sequence of latent variables $\{x_t\}_{t=1}^T$. A neural network is then trained to approximate the reverse process, progressively denoising a sample from pure Gaussian noise back into a realistic data point.

Denoising diffusion probabilistic models. Denoising Diffusion Probabilistic Models (DDPM) [14] define a forward noising process $q(x_t|x_{t-1})$ with a variance schedule $\{\beta_t\}_{t=1}^T$, where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. At an arbitrary timestep t, the closed form of the noising process is

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

The generative task is to learn the reverse process $p_\theta(x_{t-1}|x_t)$ such that a sample from $x_T \sim \mathcal{N}(0, I)$ can be gradually denoised to yield $x_0 \sim p_{\text{data}}$. In practice, this reverse transition is parameterized by a neural network $\epsilon_\theta(x_t, t)$ that predicts the noise, leading to

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}\Big(x_{t-1}; \tfrac{1}{\sqrt{\alpha_t}}\Big(x_t - \tfrac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\, \epsilon_\theta(x_t, t)\Big), \sigma_t^2 I\Big), \tag{3}$$

where σ_t^2 can be fixed or learned. Training is performed with the denoising objective

$$\mathcal{L}_{\text{denoise}}(\theta) = \mathbb{E}_{x_0, \epsilon, t}\big[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\big], \tag{4}$$

which corresponds to score matching [16], since $\epsilon_\theta(x_t, t)$ approximates the score function $-\sigma_t \nabla_{x_t} \log p(x_t)$. Moreover, by reparameterization one can directly obtain an estimate of the clean sample x_0 at timestep t as

$$\hat{x}_0(x_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\Big(x_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta(x_t, t)\Big), \tag{5}$$

which provides an explicit reconstruction of the data from noisy inputs and plays a key role in both DDPM sampling and extensions such as DDIM.

Denoising diffusion implicit models. Denoising Diffusion Implicit Models (DDIM) [42] build upon DDPM but modify the formulation to allow for a deterministic, non-Markovian sampling procedure that substantially accelerates generation. Instead of requiring T iterative reverse steps, DDIM introduces a reparameterized reverse process where the current latent x_t can be deterministically mapped to x_{t-1} using both the predicted clean image $\hat{x}_0(x_t)$ and the predicted noise $\epsilon_\theta(x_t, t)$. Specifically, the reverse update is

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, \hat{x}_0(x_t) + \sqrt{1 - \bar{\alpha}_{t-1}}\, \epsilon_\theta(x_t, t). \tag{6}$$

This deterministic formulation allows skipping intermediate steps in the reverse trajectory without retraining the model, leading to fast sampling while preserving high generative quality. DDIM thus serves as a practical alternative to DDPM and is widely adopted in applications where efficient and scalable generation is crucial.

B. Details of the multi-view diffusion model

Implementation. Our baseline model is CAT3D [12], a multi-view extension of Stable Diffusion 2.1 [38]. CAT3D adapts the latent text-to-image diffusion framework by inflating the 2D self-attention layers into 3D self-attention, enabling interactions across different views. Although the official implementation and model weights of CAT3D are not publicly available, we adopt the reproduction provided by MVGenMaster [6], which faithfully replicates CAT3D's training and evaluation pipeline.

Network architecture. The underlying architecture consists of three downsampling blocks, one mid-block, and three upsampling blocks. Each downsampling block contains two layers, the mid-block contains one layer, and each upsampling block contains three layers. Each layer comprises a spatial convolution followed by a self-attention module.

In CAT3D, standard self-attention layers are replaced with inflated 3D self-attention layers to capture inter-view

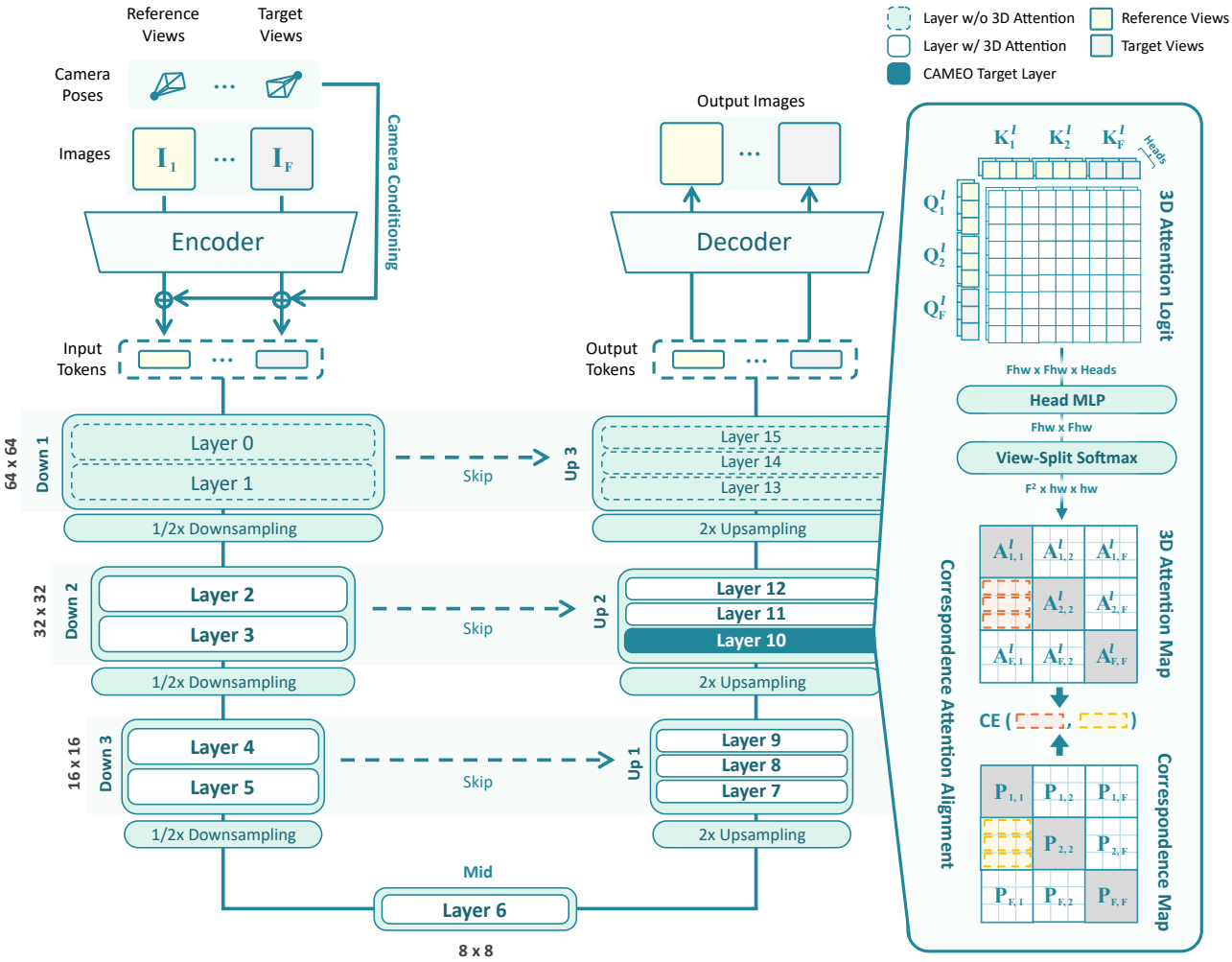

Figure 6. **Model architecture of CAT3D [12] with our proposed CAMEO framework.** While maintaining the original architecture, CAMEO introduces an additional correspondence-attention alignment loss, L_{CAMEO}, at the target layer (Layer 10) to supervise the attention map. Note that the visibility mask $\mathbf{M}_{i,j}$, which filters out occluded or unreliable correspondences via 3D cycle consistency, is omitted in this visualization for simplicity.

dependencies. This 3D attention is applied in all blocks except the first and last (*i.e.*, it is implemented in downsampling blocks 2 & 3, the mid-block, and upsampling blocks 1 & 2). In total, there are 11 inflated 3D self-attention layers used in our analysis.

The input images of resolution 512×512 are encoded by the VAE encoder into latent features of size 64×64. Gaussian noise is added to the target latents for generation, while the reference latents remain unchanged. To form the conditioning latent, we first compute the Plücker ray embedding [52], which encodes per-pixel camera rays, and concatenate it with a binary visibility mask indicating the reference images. This conditioning signal is then passed through a shallow convolutional network to match the dimensionality of the image latents. Finally, the conditioning latents are added to the image latents, producing the multi-view input representation for the diffusion U-Net.

Each downsampling block reduces the spatial resolution by a factor of 2, producing feature maps of size 32×32, 16×16, and 8×8, respectively. The mid-block operates at the lowest resolution of 8×8. The upsampling blocks then progressively restore the spatial resolution back to 16×16, 32×32, and 64×64. Finally, the latent is passed through the VAE decoder to reconstruct the full-resolution image of size 512×512.

C. Detailed analysis

In Sec. C.1, we provide the detailed qualitative analysis on U-Net [12] and DiT-based [26] model. In Sec. C.2, we provide evaluation details to measure correspondence precision

following Probe3D [11], and report the detailed analysis on CAT3D [12]. In Sec. C.3, we conduct a fine-grained analysis on the DiT-based [26] multi-view diffusion model.

C.1. Qualitative analysis

Layer-wise behavior. As shown in Fig. 7, the attention maps in deep layers $l = 10 - 12$ of CAT3D [12] consistently attends to geometrically corresponding points relative to the query points, while earlier layers ($l = 2 - 6$) do not. Layers $l = 7 - 9$ exhibit similar behavior, but the corresponding points are sparse and noisy. Similarly, Fig. 9 shows that the attention map in deep layers $l = 32 - 36$ of DiT-based model also attend to the geometrically corresponding region. This layer-wise behavior demonstrates that attention maps in deep layers of multi-view diffusion models learn to capture geometric correspondence.

Perturbation analysis. Following [1, 21], we perturb the 3D self-attention maps by enforcing them to an identity mapping, where each query token attends exclusively to itself. The generation results in Fig. 8 demonstrate that perturbing deep layers $l = 10 - 12$ of CAT3D [12] causes severe structural fragmentation and color distortion. Notably, perturbing early layers has a negligible impact on the generated images. DiT-based model exhibits consistent behavior: as shown in Fig. 10, perturbing the deep layer $l = 32$ leads to severely degraded generation, while perturbing earlier layers leaves the output largely unaffected. This implies that the geometric correspondence captured in deep layers is fundamental to maintaining the structural consistency of the generated views.

C.2. Correspondence estimation

Dataset. We evaluate geometric correspondence on the NAVI Dataset [17], which consists of 36 object scans with the ground-truth 3D geometry. The dataset provides high-quality assets, including intrinsics, extrinsics, depth maps, and object masks, enabling the inference of dense pixel-level correspondences. While Probe3D [11] utilized the *wild* split (same object, different backgrounds) of the dataset, we employ the *multiview* split to align with the Novel View Synthesis (NVS) setting, where the background remains consistent across views. To construct the evaluation set, we subsample 25% of the object views and select image pairs where the relative rotation θ of the destination view is within 120°. This results in a total of 245 source-destination pairs.

Procedure. Given two images $\mathbf{I}_1$ and $\mathbf{I}_2$ of the same scene, we aim to identify pixel pairs corresponding to the same 3D surface point. We first extract a grid of feature descriptors from each image using the corresponding backbone, and resize this grid to a spatial resolution of 128×128. Correspondences are then estimated by matching all feature descriptors from $\mathbf{I}_1$ to its nearest neighbor in $\mathbf{I}_2$ based on a predefined distance function. To mitigate inaccurate matches, we apply Lowe's ratio test [30], filtering for unique matches by comparing the distances to the first and second nearest neighbors. For each feature token p, let q_0 and q_1 be the first and second nearest neighbors, respectively. We compute the ratio r as

$$r = 1 - \frac{D(p, q_0)}{D(p, q_1)}, \tag{7}$$

where $D(x, y)$ denotes the distance between two features: cosine distance for feature-based descriptors, and ℓ_2 distance in the 3D coordinate space for pointmap representations. We retain the top 1000 matches with the highest ratio r for evaluation.

Evaluation. We evaluate correspondence precision using the 3D Euclidean distance between the reprojected points. Given a correspondence pair, we back-project both image points into a common 3D space using ground-truth depth and camera parameters. A match is deemed correct if the 3D distance between the two points is below a threshold of $\rho = 2\text{cm}$. Note that we denote this metric as *Precision* instead of *Recall* (as used in Probe3D [11]) to accurately reflect our protocol of evaluating a fixed number of top-1,000 candidates.

In Fig. 4b, we report the per-bin Precision@2cm averaged over samples in each angular bin. In the following section, we provide the detailed setup for correspondence extraction in CAT3D [12], SD2.1 [38], DINOv3 [41], and VGGT pointmap [48], Dense SIFT [27].

- **CAT3D.** For the analysis, we evaluate CAT3D [12] trained on the object-centric dataset CO3D [37]. We consider $\mathbf{I}_1$ and $\mathbf{I}_2$ as target and reference views. And we perform inference without camera conditioning and extract correspondences in a noise-free setting, *i.e.*, without injecting additional diffusion noise. Query $\mathbf{Q}_1^l$ and key $\mathbf{K}_2^l$ descriptors are extracted at each layer l (32×32 to 8×8 resolution).
- **SD2.1.** We aim to analyze the correspondence in the attention map before finetuning. Therefore, using CAT3D [12] architecture, we initialize the model weights with SD2.1 and measure the correspondence in the attention map.
- **DINOv3.** DINOv3 [41] is a state-of-the-art visual foundation model used across many downstream tasks, such as image retrieval, semantic segmentation, and dense matching. It is known for producing reliable patch-level matches. We extract patch embeddings on a 32×32 grid and compute cosine similarity between patches across views to identify correspondences. We report the results for both ViT-B/16 and ViT-L/16 variants in Tab. 10.
- **VGGT pointmap.** We measure geometric correspondence based on ℓ_2-norm nearest neighbor search. Since the pointmap explicitly encodes 3D coordinates at a res-

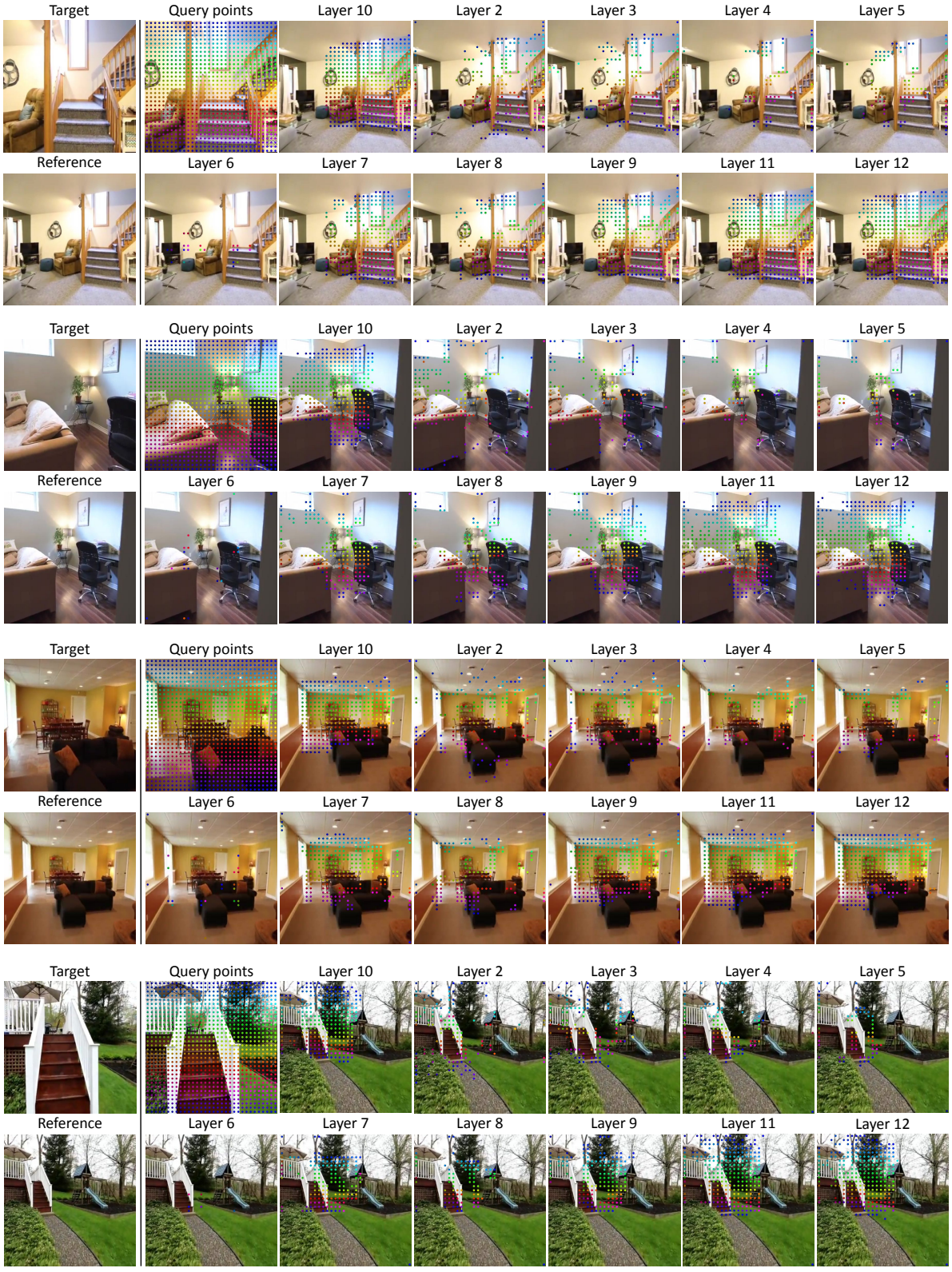

Figure 7. **Layer-wise attention map visualization of CAT3D [12].** For each query point on the target image, model's maximum attending point in the reference image is marked with the same color as the query point.

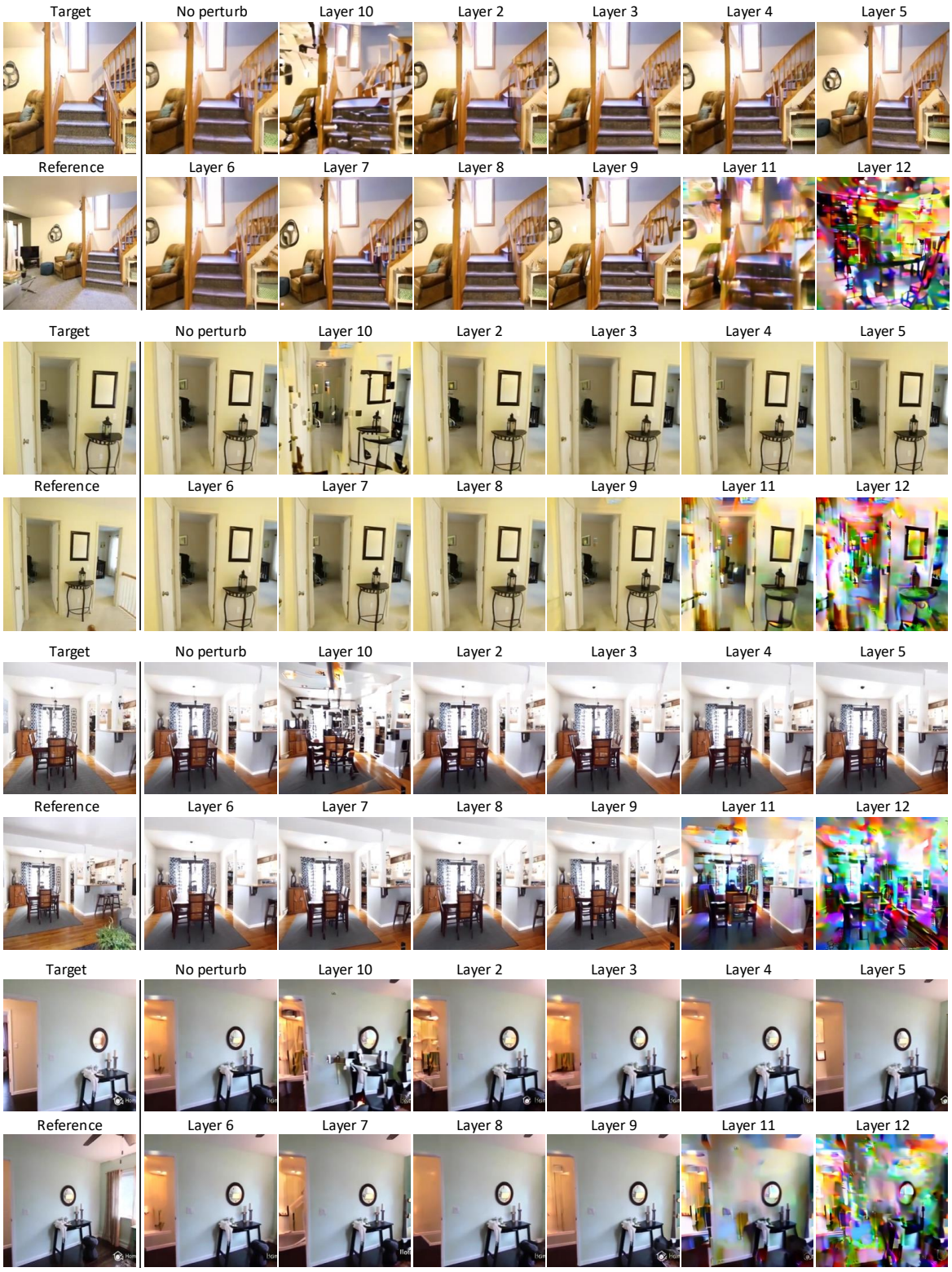

Figure 8. **Layer-wise perturbation results of CAT3D [12]** Perturbing earlier layers barely changes generation quality, while perturbing deep layers collapses geometric consistency and severely degrades quality.

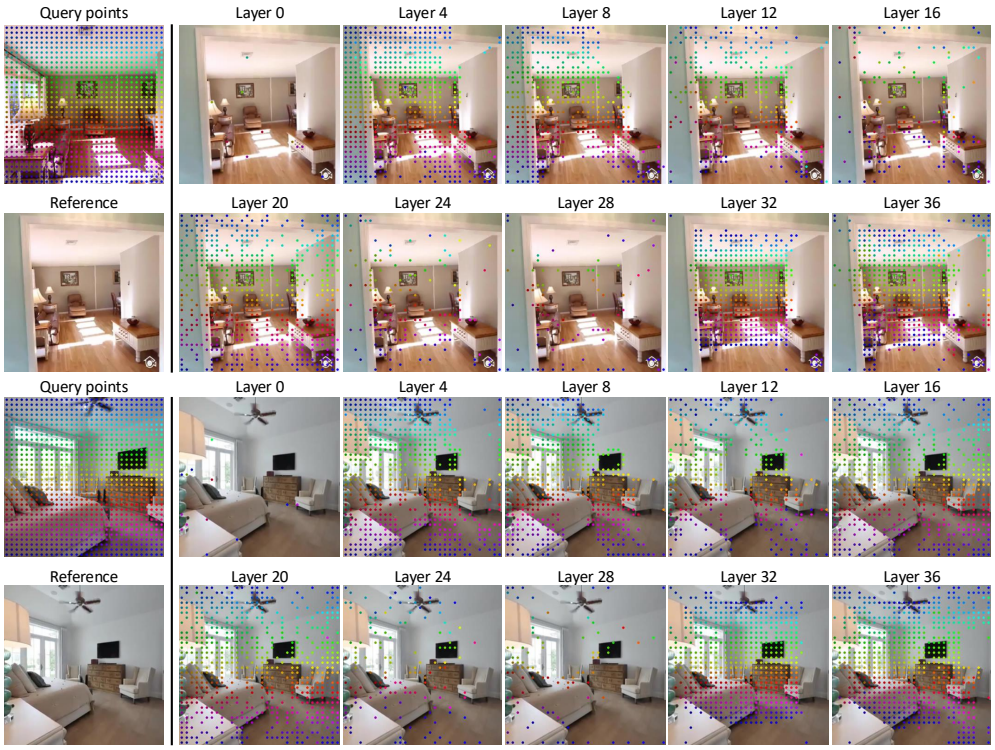

Figure 9. **Layer-wise attention map visualization of DiT-based [26] model.** For each query point on the target image, the model's maximum attending point in the reference image is marked with the same color as the query point.

Figure 10. **Layer-wise perturbation results of DiT-based [26] model.** Perturbing earlier layers barely changes generation quality, while perturbing deep layers collapses geometric consistency and severely degrades quality.

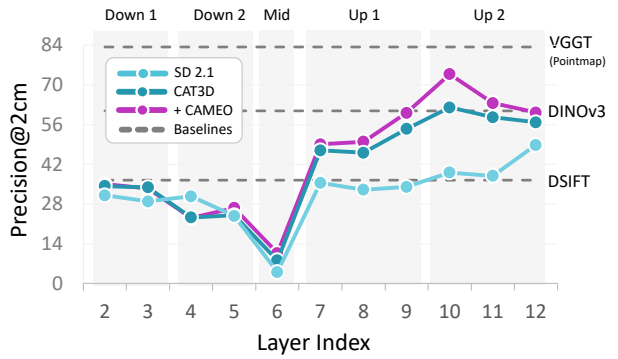

Figure 11. **Analysis of geometric correspondence in all attention layers of the multi-view diffusion model [12].** Correspondence precision across all attention layers ($l = 2-12$), with other baselines [27, 38, 41, 48].

olution of 518 × 518, we use distance-based matching rather than cosine similarity.

- **Dense SIFT.** For Dense SIFT, we follow the SIFT Flow [27] pipeline and compute dense SIFT descriptors on a multi-scale pyramid. We aggregate the descriptors onto a 128 × 128 grid and apply normalization following RootSIFT [3]. The resulting descriptors are used as feature tokens for cosine-distance matching, as in the other methods.

Correspondence precision. In Fig. 11, we present layer-wise correspondence precision across the U-Net architecture, with detailed quantitative results in Tab. 10. In the downsampling blocks ($l = 2-5$), SD2.1 [38] and CAT3D [12] exhibit comparable performance, both showing a sharp decline at the bottleneck ($l = 6$) due to spatial compression. A clear divergence emerges in the deep layers ($l = 7-12$), where CAMEO demonstrates improved consistency over SD2.1 [38]. Among all layers, $l = 10$ shows the strongest correspondence, validating our analysis in Sec. 3.2 that $l = 10$ encodes the geometric correspondence.

With alignment, CAMEO consistently achieves higher correspondence precision than CAT3D [12] across the upsampling blocks. Notably, even though our supervision is applied solely to the single target layer ($l = 10$), performance gains are observed throughout the neighboring upsampling layers ($l = 7-12$). This demonstrates that aligning one layer is sufficient to guide the model toward learning precise correspondences.

C.3. Analysis on DiT-based model

In this section, we provide a fine-grained analysis of the DiT-based [26] multi-view diffusion model to investigate its 3D self-attention map. Following the evaluation protocol in Sec. 3.2, we measure the geometric correspondence precision across all layers of the DiT model. As detailed in Fig. 12, we observe that layer $l = 32$ exhibits the highest precision. The findings in this section are consistent with those in the U-Net-based model (in Sec. 3.2), where geometric correspondence emerges most strongly in the deep attention layers.

Emergence of geometric correspondence in deep attention layers. As shown in Fig. 12, we report the geometric correspondence precision averaged over all viewpoint rotations for each layer in the DiT model. Geometric correspondence emerges most strongly in the deep layers around $l = 32$–36, where the attention maps develop significantly stronger correspondence compared to early layers.

Geometric correspondence improves throughout training. In Fig. 12, we plot the geometric correspondence precision at $l = 32$ against PSNR across training iterations of the DiT model. We observe that both correspondence precision and PSNR increase monotonically as training progresses, demonstrating that the model progressively learns to encode more accurate geometric correspondence in its attention maps. This positive correlation suggests that correspondence in the deep DiT layers is crucial for view-consistent generation, in line with our perturbation analysis in Fig. 8.

Denoising objectives provide limited correspondence supervision. Although the deep attention layers of the DiT model capture geometric correspondence and their precision correlates with generation quality, a substantial performance gap remains compared to VGGT. As shown in Fig. 12, the DiT model's attention maps exhibit significantly lower correspondence precision than VGGT. More critically, the DiT model's attention maps struggle to capture accurate geometric correspondence under large viewpoint rotations ($\theta = 60°$–$120°$), while VGGT maintains relatively robust performance across all rotation ranges. This suggests that the standard denoising objective alone is insufficient for the model to learn accurate geometric correspondence.

D. Implementation details

In Sec. D.1, we provide the evaluation settings. In Sec. D.2, we explain the detailed method to obtain correspondences from pointmaps, then evaluate the accuracy and efficiency. In Sec. D.3, we provide implementation details of a DiT-based [26] and a state-of-the-art multi-view diffusion model [6].

D.1. Evaluation settings

In Sec. 4.2, we evaluate CAMEO against baselines [12, 54] on the RealEstate10K [56], CO3D [37], and DTU [18] datasets. In the evaluation, to assess the model's robustness in maintaining view consistency under challenging scenarios, we curate a hard evaluation set. Specifically, for RealEstate10K [56] and CO3D [37], we sample images that exhibit large viewpoint changes. While this aggressive

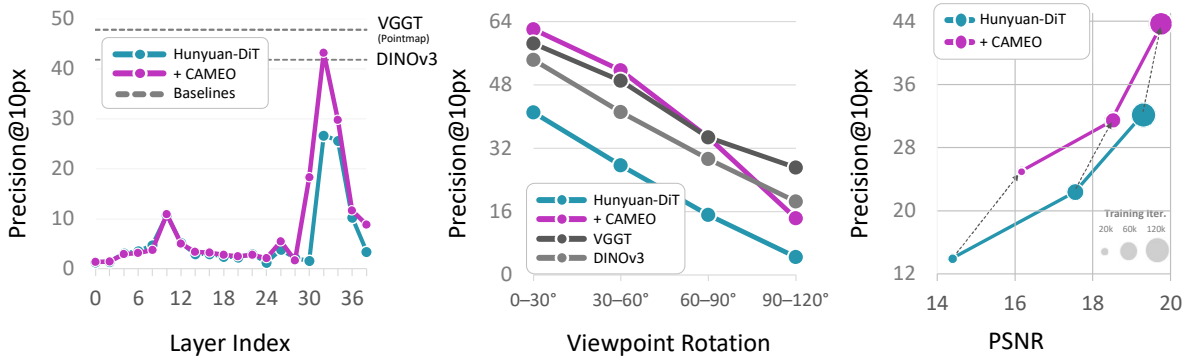

Figure 12. **Analysis of geometric correspondence in attention maps of DiT-based multi-view diffusion model [26].** Leftmost: correspondence precision across all layer indices, with baselines [41, 48]. Middle: correspondence precision of layer $l = 32$ across viewpoint rotations. Rightmost: correspondence precision of layer $l = 32$ improves during training alongside PSNR.

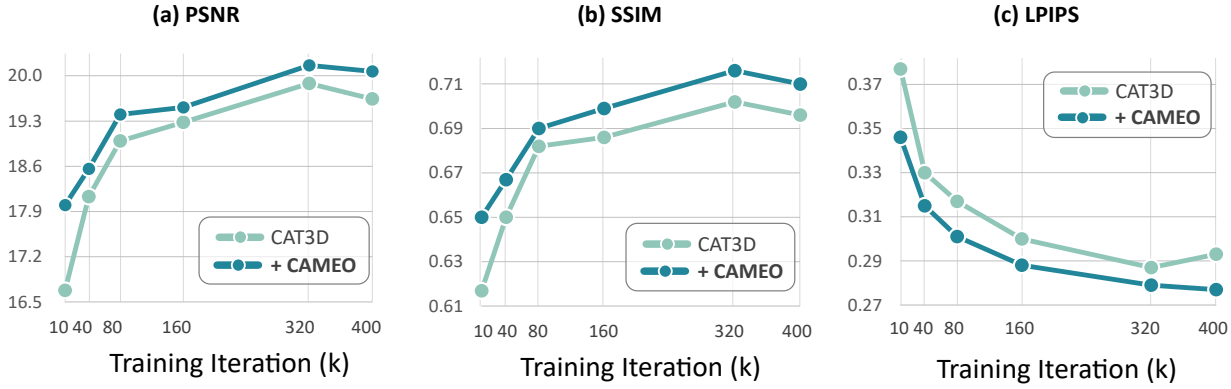

Figure 13. **The relative improvements of CAMEO over CAT3D [12] on RealEstate10K [56] dataset.**

sampling strategy naturally yields lower quantitative scores across all evaluated models (as seen in Tab. 1), it is essential for identifying failure cases in geometric consistency and rigorously assessing robustness under challenging scenarios.

D.2. Correspondence from pointmap

Method. DUSt3R [49] introduces an algorithm to establish pixel correspondences between two images through nearest neighbor search in 3D pointmap space. Specifically, for an image $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$, a pointmap $\mathbf{X}_i \in \mathbb{R}^{H\times W\times 3}$ represents the 3D coordinates of each pixel. Given an image pair $(\mathbf{I}_i, \mathbf{I}_j)$ where $i \neq j$, correspondences are established by computing mutual nearest neighbors between pixel locations $\mathbf{y}_i$ and $\mathbf{y}_j$:

$$\{(\mathbf{y}_i, \mathbf{y}_j) \mid \mathbf{y}_j = \mathrm{NN}^{i,j}(\mathbf{y}_i) \text{ and } \mathbf{y}_i = \mathrm{NN}^{j,i}(\mathbf{y}_j)\}, \quad (8)$$

where $\mathrm{NN}^{i,j}(\mathbf{y}_i) := \underset{k\in\{1,\dots,HW\}}{\arg\min} \|\mathbf{X}_i(\mathbf{y}_i) - \mathbf{X}_j(k)\|_2$ denotes the nearest neighbor of pixel $\mathbf{y}_i$ in view j within 3D space. We extend this algorithm to token-level resolution by downsampling the pointmaps, and introduce a cycle consistency threshold τ to replace the exact mutual matching criterion.

Accuracy. We employ VGGT [48] to obtain pointmaps, as it reports superior performance compared to previous geometry prediction models, such as DUSt3R [49] and MASt3R [23]. As illustrated in Fig. 4a, VGGT pointmaps achieve the correspondence precision of 83.32, significantly outperforming DINOv2 at 60.84 and Dense SIFT [27] at 36.43.

Efficiency. VGGT [48] employs a simple feed-forward approach that build pointmaps in only 0.2 seconds. Although tracking is an alternative method for producing dense correspondences between images, dense tracking is computa-

Table 7. **Runtime and peak GPU memory usage across different numbers of input frames.** Runtime is measured in seconds, and GPU memory usage is reported in gigabytes.

Input Frames		1	2	4	8	10	20	50
Time (s)	Pointmaps	**0.13**	**0.16**	**0.27**	**0.52**	**0.68**	**1.73**	**5.46**
	Tracking	1.44	2.95	6.30	14.10	18.91	53.10	304.58
Mem. (GB)	Pointmaps	**2.1**	**2.4**	**3.1**	**4.5**	**5.3**	**8.8**	**19.5**
	Tracking	2.3	2.7	3.4	5.0	5.7	9.6	21.1

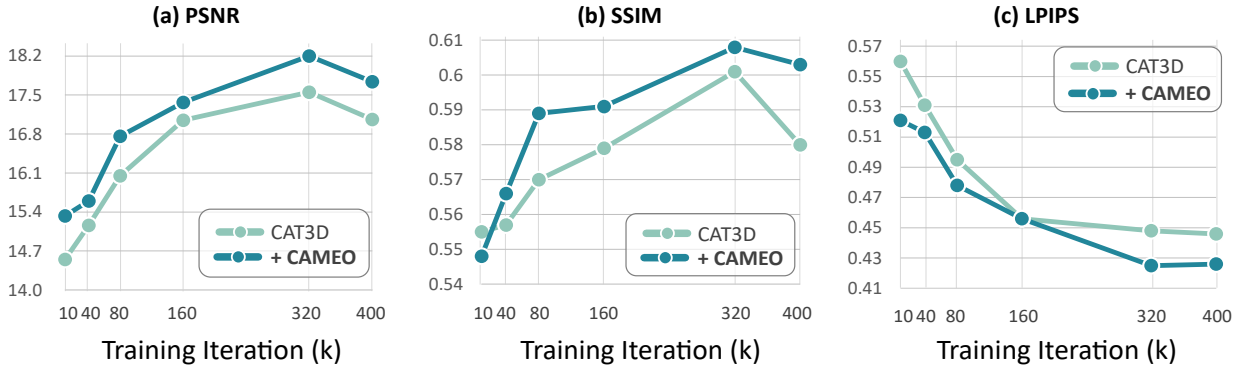

Figure 14. **The relative improvements of CAMEO over CAT3D [12] on CO3D [37] dataset.**

tionally prohibitive. Moreover, tracking requires F separate inferences to supervise the 3D self-attention map, where F is the total number of views. We evaluate inference runtime and peak GPU memory for obtaining pointmaps and tracks from VGGT [48] with varying numbers of input frames. Measurements are conducted using a single NVIDIA RTX A6000 GPU with images at 336×518 resolution. For tracking, we use 1024 query points, equivalent to the total number of query tokens in the attention map at layer $l = 10$ of CAT3D [12]. As shown in Tab. 7, pointmap inference is significantly faster and more memory-efficient than tracking.

D.3. Other architectures

State-of-the-art multi-view diffusion model. We implement MVGenMaster [6] based on the officially released code. We use an off-the-shelf geometry model [48] to obtain depth maps for geometric conditions. We initialize the model with Stable Diffusion 2.1 [38] weights, and train the model on the RealEstate10K [56] dataset with a batch size of 3. Other training and evaluation details are identical to those in the main model experiment.

DiT-based multi-view diffusion model. Following Matrix3D [31], we implement a multi-view diffusion model based on a pre-trained text-to-image diffusion transformer (DiT) [26]. However, instead of using an external transformer encoder to embed conditional inputs as in Matrix3D [31], we adopt a simpler approach by inflating the self-attention to 3D self-attention, similar to our baseline [12]. Specifically, we concatenate the query, key, and value matrices of each self-attention layer, and omit the cross-attention layer and text encoder. Following Matrix3D [31], we employ Rotary Positional Embedding (RoPE) [43] to encode each token's position and absolute sinusoidal positional encoding [10] to encode the viewpoint index. We use Plücker rays to represent camera poses and add the camera pose embeddings as residuals. The model is initialized with Hunyuan-DiT [26] weights and trained on the RealEstate10K [56] dataset with a batch size of 4. All other training and evaluation details remain identical to those of the main model.

E. Discussion

We analyze the effect of CAMEO by measuring correspondence precision and visualizing the attention maps of CAT3D [12] trained with and without CAMEO.

Correspondence accuracy. Fig. 4 presents two observations. In Figs. 4a and 4b, CAMEO increases attention correspondence precision across all viewpoint rotations, even surpassing feature matching in DINOv3 [41]. As shown in Fig. 4c, CAMEO can push the baseline [12] to achieve both higher correspondence precision and PSNR at the same iterations, indicating faster learning of geometric correspondence and earlier gains in generation quality. In Fig. 12, we observe consistent improvements in the DiT-based model [26]: CAMEO substantially boosts correspondence precision at the deep layer $l = 32$, surpassing DINOv3 and narrowing the gap to VGGT [48] across all viewpoint rotations, while simultaneously achieving higher PSNR at the same training iterations.

Qualitative analysis. In Fig. 15, CAT3D [12] produces a distorted handrail, while CAMEO preserves the correct shape with fine detail and accuracy. The attention maps reveal the mechanism behind this performance gap. For CAT3D [12], query points on the handrail fail to attend to the handrail region in the reference image. In contrast, CAMEO correctly attends to the corresponding handrail region, resulting in precise geometric reconstruction with structural details. This demonstrates that CAMEO successfully guides the model to learn more accurate geometric correspondences, which directly improves novel view synthesis performance.

F. Ablation studies

Feature costmap. Motivated by DIFT [46], which demonstrates that intermediate features from Stable Diffusion

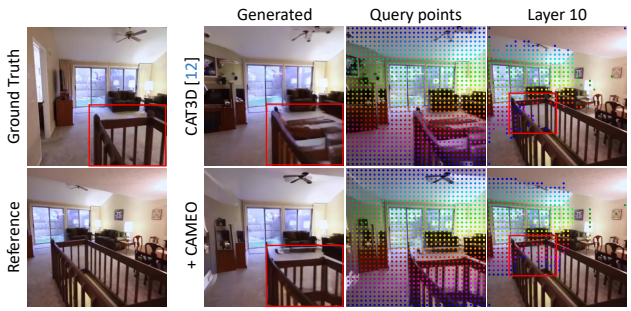

Figure 15. **Correspondence analysis in $l = 10$.** In CAT3D [12], pink query points on the handrail fail to attend to their geometric counterparts in the reference, whereas CAMEO succeeds. As a result, the handrail is accurately generated only in CAMEO.

Table 8. **Ablation study for supervision target.** Evaluated on RealEstate10K [56] at 40k iterations.

Supervision Targets	PSNR↑	SSIM↑	LPIPS↓
Feature Cost	18.17	0.650	0.333
Attention	**18.42**	**0.662**	**0.323**

Table 9. **CAMEO with REPA.**

Method	Iter.	PSNR↑	SSIM↑	LPIPS↓
CAT3D	10k	16.68	0.617	0.377
w/ CAMEO		**18.00**	**0.650**	**0.346**
w/ CAMEO + REPA		17.83	0.647	0.350
CAT3D	40k	18.13	0.650	0.330
w/ CAMEO		**18.56**	**0.666**	**0.315**
w/ CAMEO + REPA		17.93	0.651	0.329

2.1 [38] encode correspondence, we investigate the impact of the supervision target. Specifically, we compare two approaches at layer $l = 10$: (1) supervising the feature similarity map derived from the layer's output features, and (2) directly supervising the attention map. We train the models with a batch size of 6, while other training and evaluation settings remain identical to the ablation experiments. Tab. 8 shows that supervising the attention map yields superior performance. This confirms that geometric correspondence is most effectively regulated directly within the attention mechanism, whereas feature similarity requires computing an additional cost map extraneous to the original architecture.

Compatibility with REPA. Combining CAMEO with REPA degrades performance (Tab. 9). Multi-view diffusion models are finetuned from pretrained models, whereas REPA is designed for 'from scratch' training. We hypothesize that once representations are established, REPA's effectiveness becomes limited. Moreover, supervising multiple layers with different objectives introduces optimization sensitivity, destabilizing training. We leave a more principled integration of CAMEO with representation alignment objectives into finetuning-based multi-view diffusion as an avenue for future work.

G. Qualitative results

We provide additional qualitative comparisons of CAMEO on both scene-level [56] and object-centric [18, 37] settings. We present qualitative comparisons against the baseline, highlighting improved geometric consistency. Figs. 17 to 19 shows qualitative examples organized by training iteration. We also provide the qualitative results of CAMEO on the state-of-the-art model [6] and DiT-based [26] model in Fig. 20 and Fig. 21.

H. 3D reconstruction

Following CAT3D [20], we also perform 3D reconstruction using the novel views generated by the model. We create camera trajectories to generate novel view images and use them to optimize 3DGS [20]. Specifically, we first run the multi-view diffusion models on 2-view settings, where the first and the last cameras are input views, and sample target camera trajectories between them evenly. The total number of views are 100 (2 input views and 98 generated views). We then optimize 3DGS with ℓ_1, SSIM loss, alongside LPIPS loss following CAT3D [12], to reconstruct 3D scenes from generated novel views.

We provide the 3D reconstruction results on DTU [18] dataset in Fig. 16. While CAT3D [12] fails to reconstruct 3D scenes, CAMEO can faithfully reconstruct scenes through 3DGS. This demonstrates that CAMEO produces view-consistent images, leading to higher-quality 3D reconstructions than the baseline [12].

I. Limitations

Our method may struggle with extreme viewpoint changes where reference and target views share minimal or no visual overlap. In such scenarios, establishing cross-view correspondence becomes inherently infeasible. Since CAMEO is designed to leverage geometric correspondences between views, its effectiveness is naturally constrained under extreme viewpoint gaps. This reflects a fundamental challenge in novel view synthesis. To address such scenarios, alternative strategies can be employed, sequentially generating intermediate views with each step conditioned on previously generated images [25, 47].

J. Future work

- **Beyond novel view synthesis.** Our method targets multi-view diffusion for novel view synthesis. Extending correspondence-aware supervision to video diffusion, 4D reconstruction, or other multi-modal tasks remains an open direction.

Figure 16. **3D reconstruction results.** We input 2 views and generate novel views to optimize 3DGS [20]. In the 3DGS rendering results, CAMEO exhibits consistent rendered images while CAT3D [12] fails.

- **Semantic correspondences.** We demonstrate that specific layers encode geometric correspondence and improve performance through geometric alignment. As an extension, semantic correspondence may be encoded in other layers, and leveraging this signal could further enhance generation quality and semantic understanding.

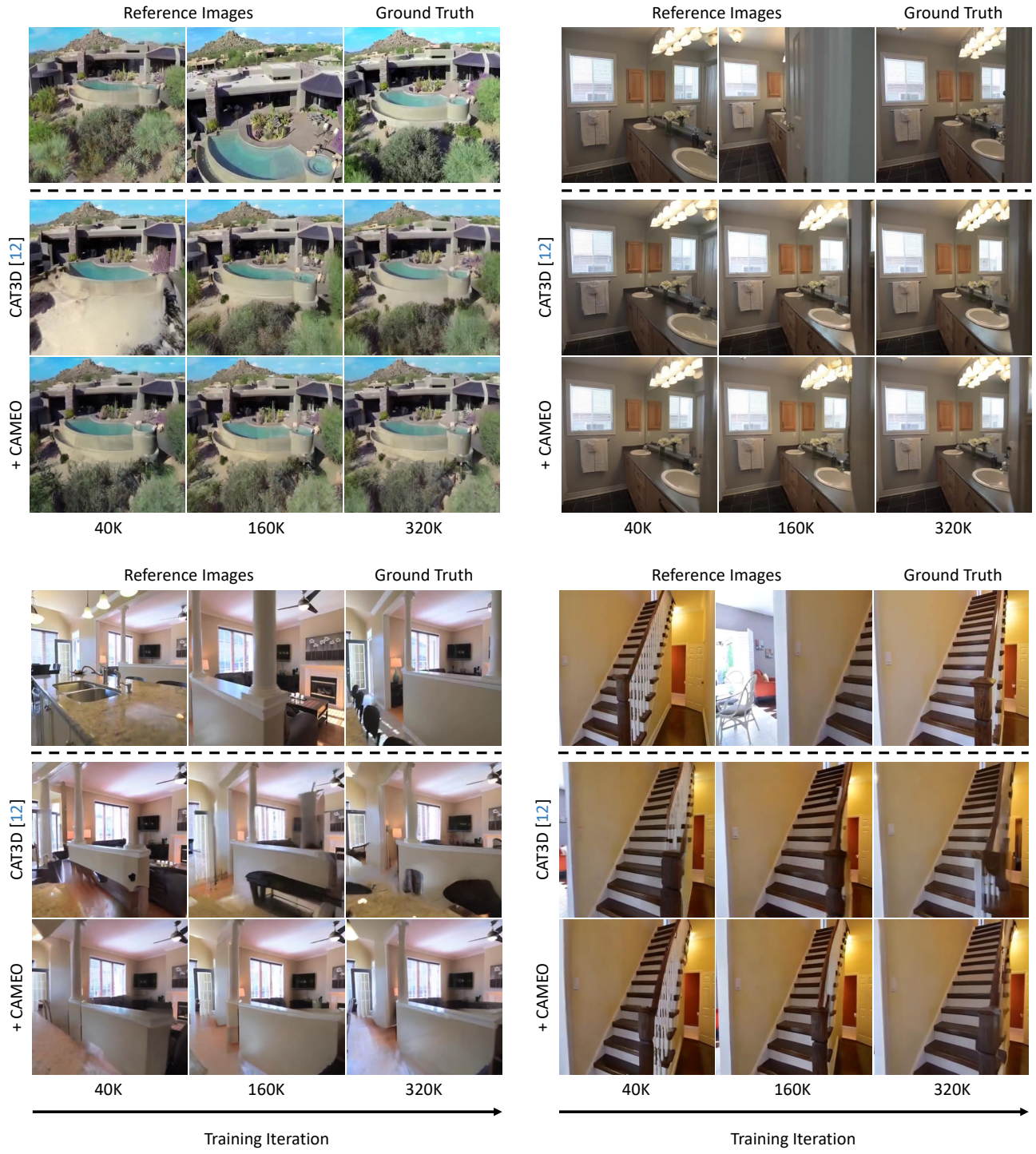

Figure 17. **Qualitative results** on RealEstate10K [56]. CAMEO improves learning efficiency while significantly enhancing geometric consistency compared to the baseline, as explicit correspondence supervision encourages faster convergence in novel view synthesis.

Figure 18. **Qualitative results** on CO3D [37]. CAMEO improves learning efficiency while significantly enhancing geometric consistency compared to the baseline, as explicit correspondence supervision encourages faster convergence in novel view synthesis.

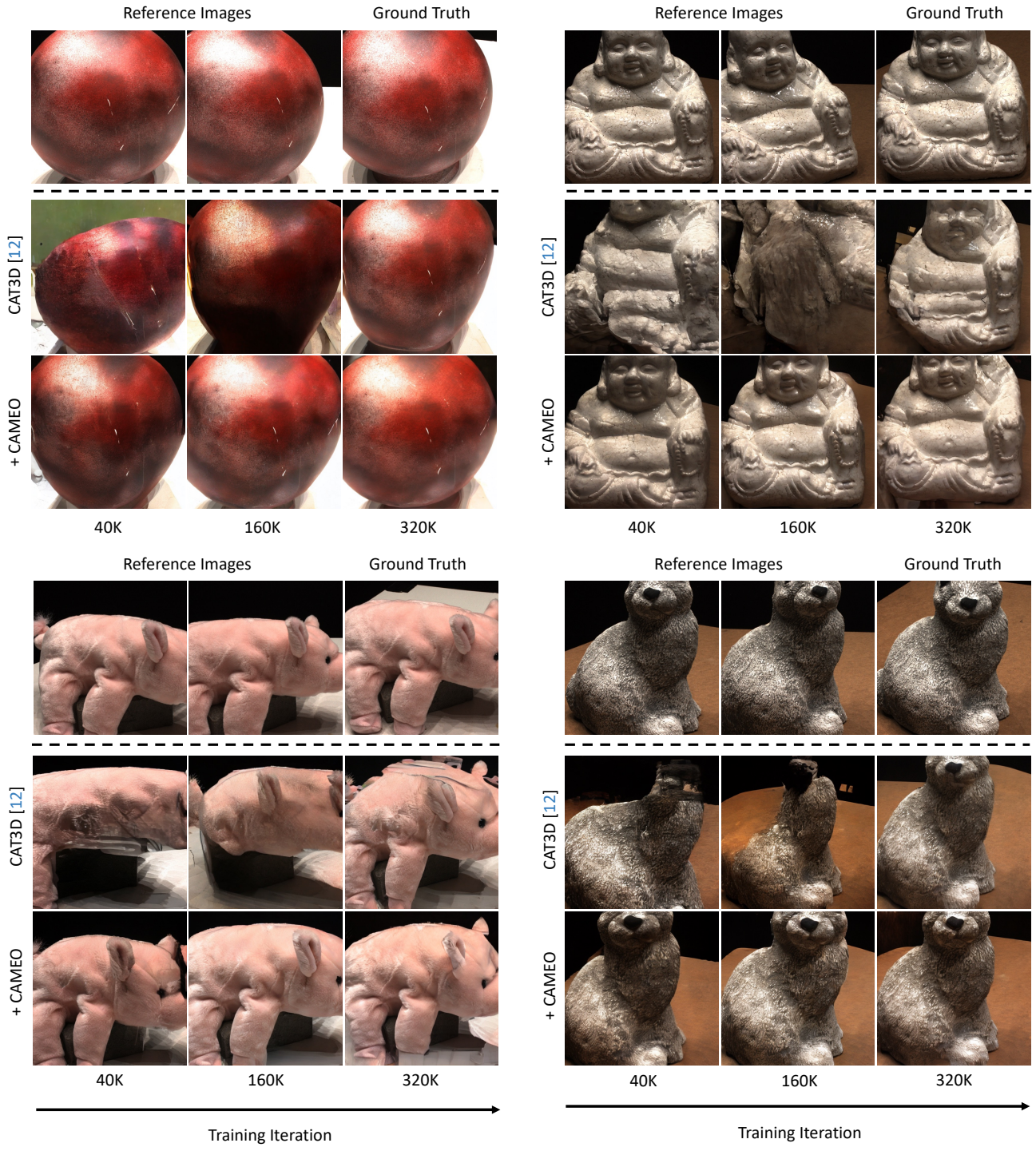

Figure 19. **Qualitative results** on DTU [18] (Out-of-domain). CAMEO improves learning efficiency while significantly enhancing geometric consistency compared to the baseline, as explicit correspondence supervision encourages faster convergence in novel view synthesis.

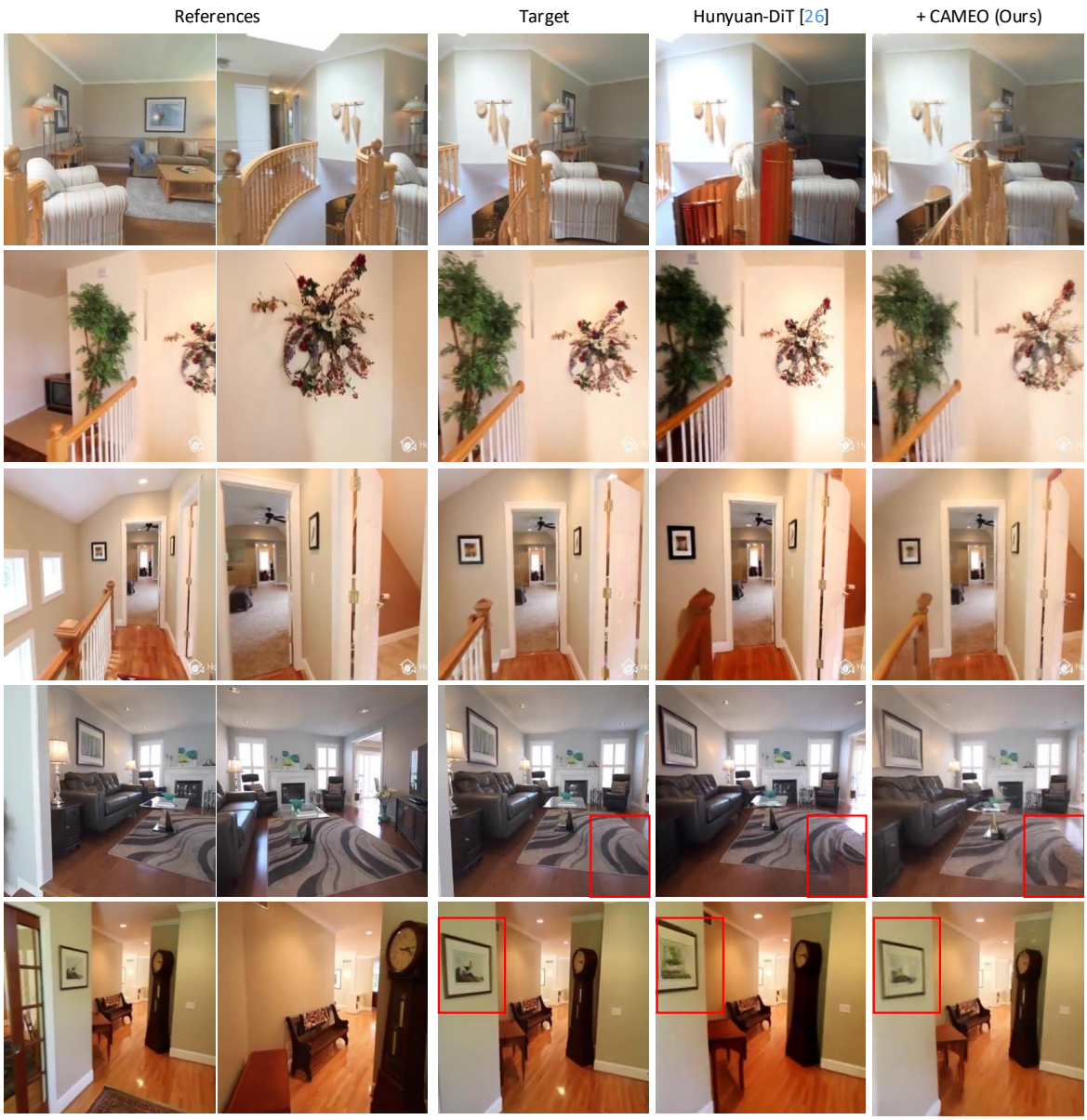

Figure 20. **Qualitative results of DiT-based model [26] on RealEstate10K [56].** CAMEO enhances geometric consistency compared to the baseline. By incorporating explicit correspondence supervision, our method encourages the model to learn and preserve accurate structural relationships across views.

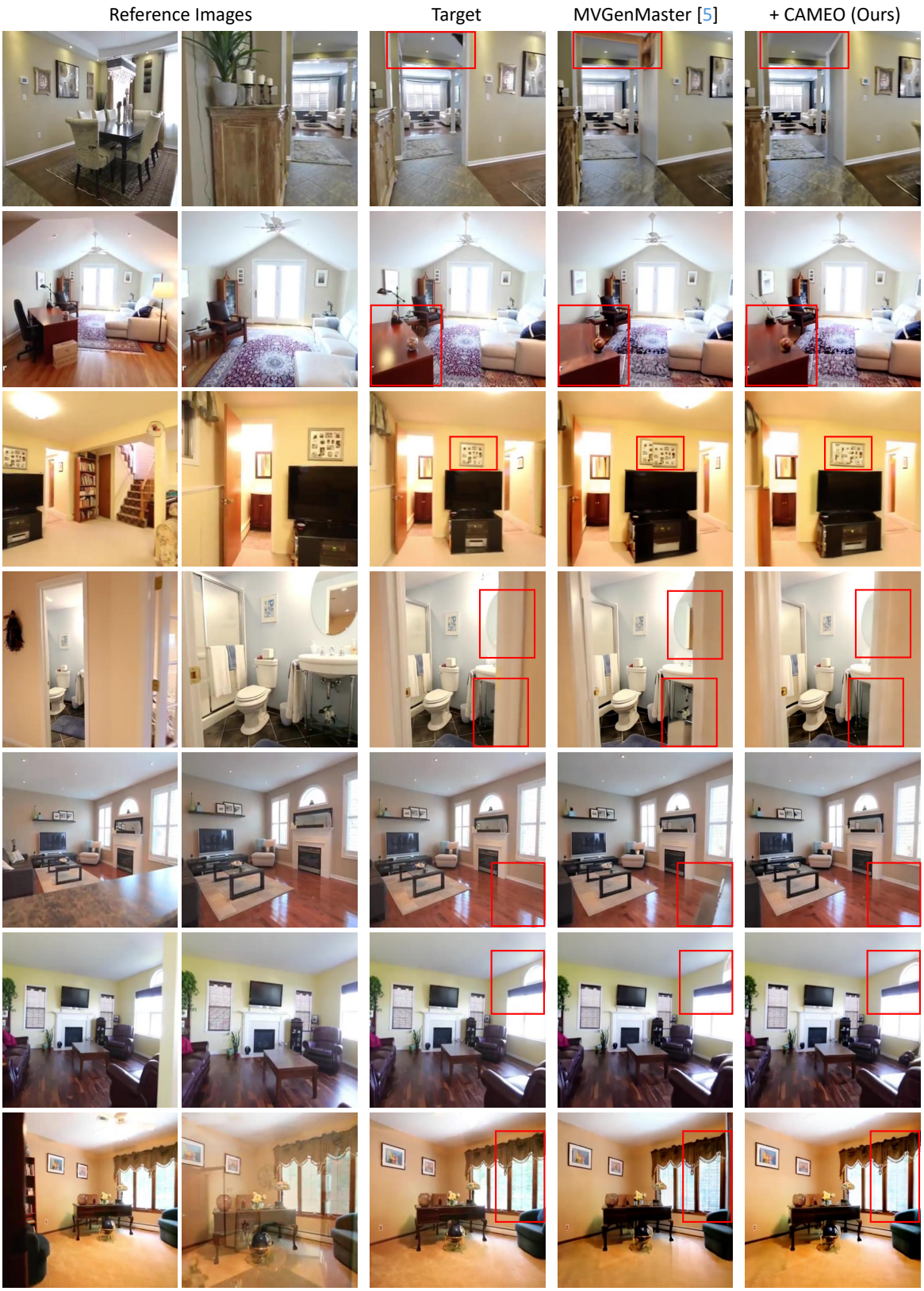

Figure 21. **Qualitative results of MVGenMaster on RealEstate10K [56].** CAMEO enhances geometric consistency compared to the baseline. By incorporating explicit correspondence supervision, our method encourages the model to learn and preserve accurate structural relationships across views.

Table 10. **Performance comparison on the NAVI [17] dataset.** We report Precision@2cm following the Probe3d [11] protocol. Best results per category are marked in **bold**. CAT3D [12] and CAMEO are trained on the Co3D [37] dataset for 320k training iterations.

Model		Feature Grid	Precision@2cm				
			Overall	**$0-30^\circ$**	**$30-60^\circ$**	**$60-90^\circ$**	**$90-120^\circ$**
VGGT Pointmaps [48]		518×518	**83.32**	**97.41**	**92.28**	**83.12**	**63.86**
Dense SIFT [27]		128×128	36.43	90.02	53.82	14.51	8.68
DINOv3 [41]	ViT-L/16	32×32	**60.84**	**95.66**	**78.36**	**50.51**	**30.22**
	ViT-B/16	32×32	56.68	94.34	74.91	44.05	26.18
SD2.1 [38]	2	32×32	31.09	52.88	39.32	23.07	17.28
	3		28.94	45.68	36.17	21.97	18.09
	4	16×16	30.75	57.41	39.66	22.46	13.55
	5		23.81	37.77	31.97	18.59	11.32
	6	8×8	4.05	1.67	4.36	5.70	3.08
	7	16×16	35.47	72.65	49.14	21.94	12.18
	8		33.04	71.89	45.55	19.48	10.33
	9		34.07	70.16	45.83	21.87	12.33
	10	32×32	39.12	84.87	56.98	21.52	10.25
	11		37.95	80.87	53.71	21.20	12.42
	12		**48.85**	**90.24**	**68.47**	**32.51**	**18.71**
CAT3D [12]	2	32×32	34.36	63.67	45.69	22.80	16.31
	3		33.90	62.19	44.60	23.31	16.13
	4	16×16	23.28	40.96	30.43	17.01	11.17
	5		24.06	41.87	32.70	18.08	9.62
	6	8×8	8.20	10.44	10.28	8.16	4.25
	7	16×16	46.95	76.76	61.23	39.89	19.57
	8		46.07	78.65	61.05	36.65	18.94
	9		54.50	84.96	69.48	45.64	27.92
	10	32×32	**62.07**	**94.33**	**79.63**	**53.73**	**30.51**
	11		58.61	92.41	77.70	48.06	26.77
	12		56.82	80.80	73.97	45.93	27.80
CAMEO	2	32×32	35.05	64.84	46.83	24.28	15.24
	3		33.53	59.32	44.21	24.29	15.66
	4	16×16	23.10	39.91	30.60	16.85	10.65
	5		26.61	45.17	35.18	20.82	11.60
	6	8×8	10.74	11.75	12.03	9.43	10.00
	7	16×16	49.08	81.05	64.56	41.74	19.23
	8		49.97	76.29	62.59	45.38	23.88
	9		60.19	86.81	71.84	55.57	35.22
	10	32×32	**73.80**	**94.88**	**84.80**	**72.92**	**48.50**
	11		63.62	93.23	79.89	57.12	33.11
	12		60.30	90.89	75.21	51.11	33.00