

Table 5: **Cross-dataset generalization.** We conduct zero-shot tests on the DTU and ACID datasets using models trained on RealEstate10K without any fine-tuning.

Method	RealEstate10K→DTU			RealEstate10K→ACID		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
pixelSplatCharatan et al. (2024)	12.89	0.382	0.560	27.64	0.830	0.160
MVSplatChen et al. (2025b)	13.94	0.473	0.385	28.15	0.841	0.147
TranSplatZhang et al. (2025)	14.93	0.531	0.326	28.17	0.842	0.146
HiSplatTang et al. (2025b)	16.05	0.671	0.277	28.66	0.850	0.137
CoDiffSplat (Ours)	16.31 (+0.26)	0.704 (+0.033)	0.269 (−0.008)	28.70 (+0.04)	0.853 (+0.003)	0.138 (+0.001)

A APPENDIX

A.1 IMPLEMENTATION DETAILS

We implement CoDiffSplat in PyTorch. Following prior baselines, all input images are resized to 256×256 , and we adopt two sparse views unless otherwise noted. We set downsampling rate $s = 4$ to obtain a 64×64 latent resolution. For semantic feature extraction, we use the DINOv3-pretrained Caron et al. (2021) ViT-Base Dosovitskiy et al. (2021). The diffusion backbone is based on TinyDiT Fang et al. (2025), an effective DiT Peebles & Xie (2023) architecture. We train the model for 300,000 iterations using one NVIDIA H800, with the AdamW optimizer Loshchilov & Hutter (2019) and a batch size of 8.

A.2 ADDITIONAL DISCUSSION

A.2.1 CROSS-DATASET GENERALIZATION

To assess zero-shot cross-dataset generalization, we evaluate on the real object-centric DTU dataset Jensen et al. (2014), selecting 16 validation scenes with four novel views each. Our proposed CoDiffSplat inherently excels at generalizing to out-of-distribution (OOD) scenes, due to the frozen ViT-based semantic embeddings with robust representational capacity. To verify this advantage, we perform cross-dataset evaluations following the protocol of MVSplat Chen et al. (2025b). Specifically, we train the model on RealEstate10K (an indoor dataset) and directly test on ACID (an outdoor dataset) and DTU (an object-centric dataset). As illustrated in Figure 5, despite the substantial differences in camera distributions and visual appearance between these datasets, CoDiffSplat consistently produces superior novel-view renderings, whereas baselines frequently exhibit artifacts and geometric inconsistencies. The discrepancy arises because the baselines rely on forcibly fitting depth distributions, whereas our conditional diffusion approach leverages high-level semantic information, mitigating the need for precise depth estimation.

Quantitative results in Table 5 demonstrate the superior cross-dataset generalization of CoDiffSplat on both DTU and ACID. On DTU, it outperforms the second-best HiSplat by **+0.26 PSNR**, **+0.033 SSIM**, and **-0.008 LPIPS**, reflecting higher fidelity and more perceptually accurate renderings. On ACID, despite HiSplat achieving the lowest LPIPS (0.137), CoDiffSplat still delivers consistent gains in distortion-based metrics (**+0.04 PSNR** and **+0.003 SSIM**). These results validate that semantic-aware diffusion provides stronger robustness than purely geometry-driven estimators, supporting our claim that semantic embeddings coupled with diffusion priors are an effective path to OOD generalization.

We attribute the overall superiority of CoDiffSplat not only to its semantic-aware design but also to the larger training scale of RealEstate10K (approximately seven times larger than ACID), which further enhances its ability to generalize. These findings suggest that training CoDiffSplat on larger and more diverse datasets could yield even stronger performance in challenging OOD scenarios.

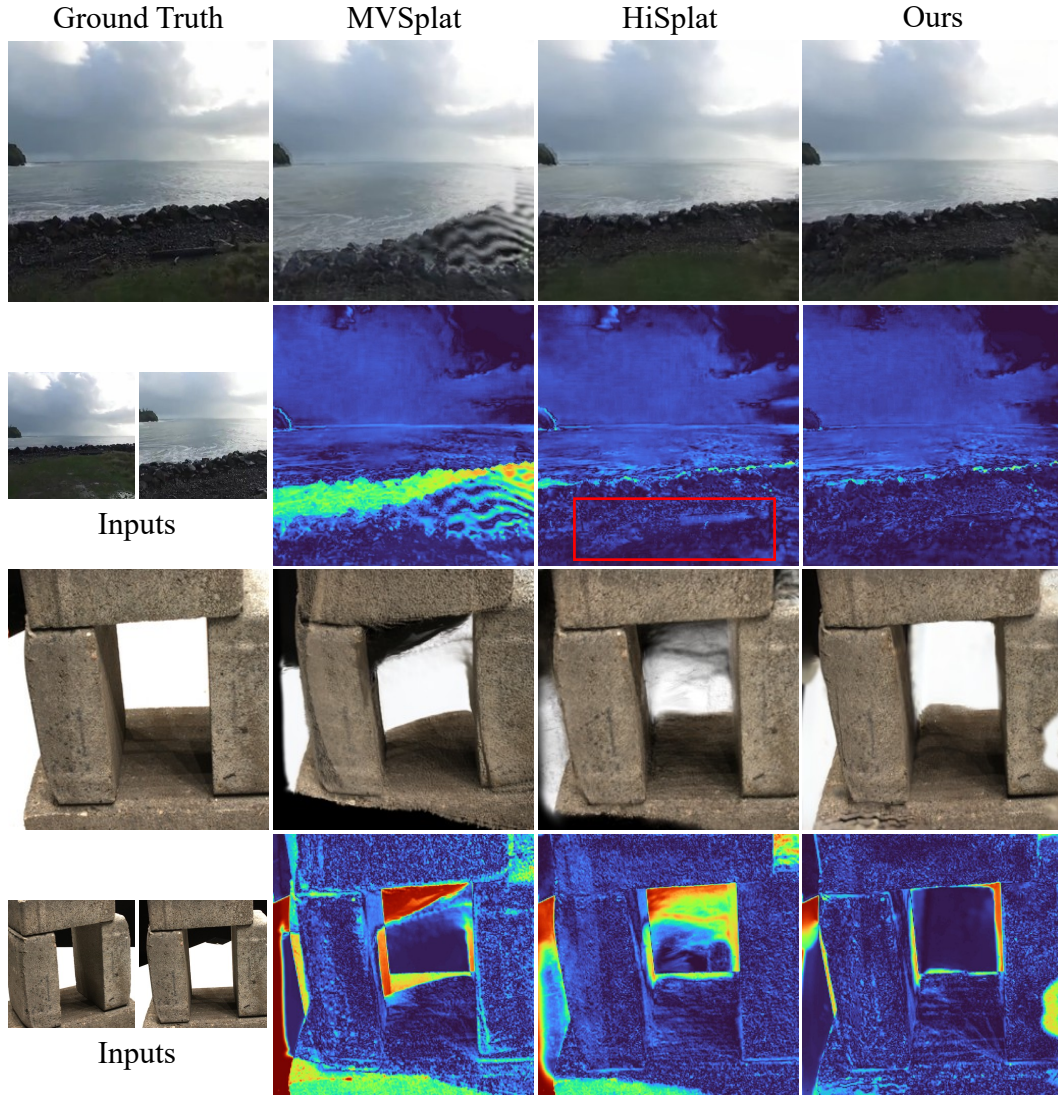
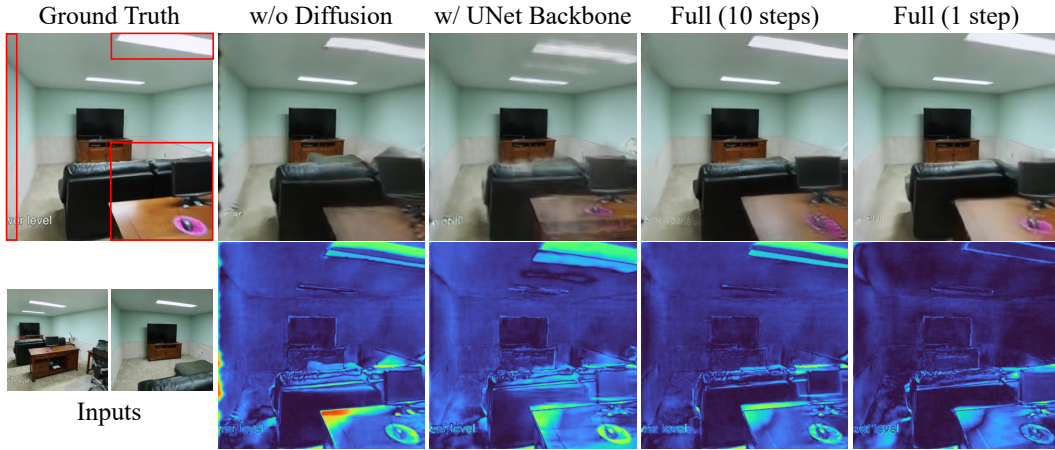
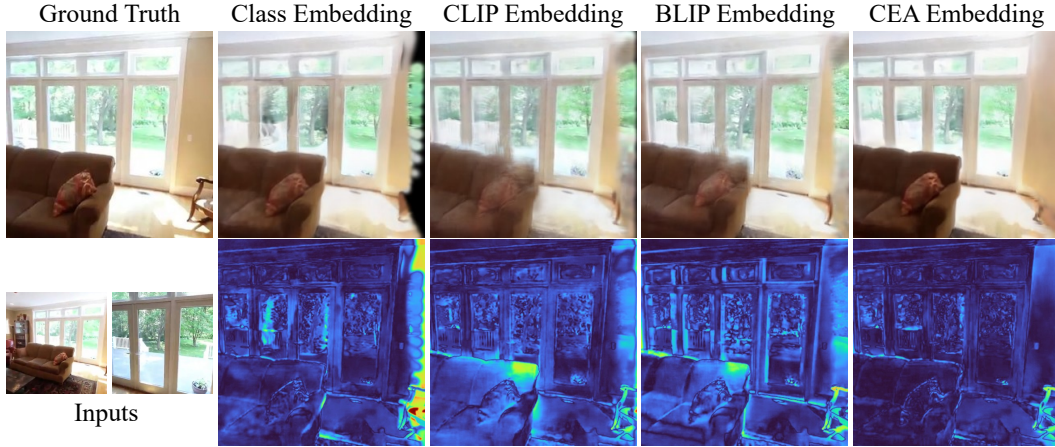


Figure 5: **Qualitative comparison of cross-dataset generalization.** Models trained on RealEstate10K are used to conduct zero-shot test on scenes from ACID (top row) and DTU (bottom row), without any fine-tuning.

Figure 6: **Qualitative ablations of Diffusion.**Figure 7: **Qualitative ablations of various embeddings.**

A.2.2 QUALITATIVE ANALYSIS OF DIFFUSION

Figure 6 provides qualitative comparisons under challenging viewpoint changes. Without diffusion, we observe severe structural distortions and missing geometry, especially near object boundaries. While a UNet backbone produces more complete structures than the non-diffusion baseline, it introduces blurry edges and artifacts (e.g., distorted lights). We attribute to the inductive bias of UNet, which imposes artificial 2D grid-structured constraints unsuited to irregular 3D Gaussian distributions. In contrast, DiT better preserves spatial irregularities of 3D Gaussians, yielding sharper details. For diffusion steps, multi-step denoising tends to oversmooth fine structures and shift their positions (e.g., lights and monitors), validating that our single-step design offers a stable yet precise refinement. These observations align with the quantitative trends in Table 3.

A.2.3 QUALITATIVE ANALYSIS OF VARIOUS EMBEDDINGS

Qualitative results shown in Figure 7 reveal that the other embeddings fail to accurately recover edge details (the right-side wall). This issue may stem from primarily focusing on larger objects and neglecting boundary information. The class embedding introduces tree-like artifacts outside the window, as it provides coarse semantic information that cannot effectively guide weak-texture alignment. Both the CLIP and BLIP embeddings struggle to capture finer details of the sofa. This

Table 6: **Quantitative comparison of 3-view Cross-Dataset Generalization.**

Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
pixelSplatCharatan et al. (2024)	12.52	0.367	0.585
MVSplatChen et al. (2025b)	14.30	0.508	0.371
CoDiffSplat (Ours)	16.44 (+2.14)	0.729 (+0.221)	0.298 (-0.073)



Figure 8: **More comparisons on RealEstate10K.** Qualitative comparison of novel view synthesis results across different methods. Red boxes highlight challenging regions where baseline methods exhibit artifacts or texture loss. In contrast, CoDiffSplat produces sharper and more structurally consistent renderings, refining object contours and details (e.g., furniture edges, patterned bedspreads).

could be attributed to their sensitivity to multi-view overlap and redundancy, which diminishes the attention to unique objects appearing only once. In contrast, the CEA embedding effectively refines and repairs the scene, demonstrating its ability to perceive uncertain or weakly constrained regions and remove redundant information from overlapping areas.

A.2.4 GENERALIZATION TO MORE INPUT VIEWS

To assess the generalizability of our model under varying numbers of sparse input views, we conduct a zero-shot evaluation by directly applying the model trained with 2-view inputs in RealEstate10K to a 3-view input setting on DTU. As reported in Table 6, our method demonstrates robust performance when evaluated with more input views.

A.2.5 MORE VISUAL COMPARISONS

We provide additional qualitative comparisons on the RealEstate10K dataset, as shown in Figure 8. PixelSplat and MVSplat frequently exhibit severe blurring and geometric distortions, while HiSplat produces sharper outputs but still fails to recover boundaries and occluded regions. In contrast, CoDiffSplat consistently generates sharper novel views, preserving furniture outlines and textile patterns.

Figure 9: **Failure case.**

A.2.6 LIMITATIONS

Our design of single-step diffusion effectively refines Gaussians and preserves geometric consistency, yet it inherently limits the capacity to hallucinate structures that are entirely absent in the inputs. Specifically, although semantic guidance allows the model to inpaint partially missing or ambiguous areas, it cannot reliably generate geometry in regions that are fully unseen and semantically uninformative, as shown in Fig. 9. Future progress may benefit from larger scene-level datasets with explicit 3D supervision, which could provide stronger priors for extending diffusion-based 3DGS beyond the observed distributions.