INFINISPLAT V2: ANALYZING AND IMPROVING FEED-FORWARD 3DGS FROM ANY VIEWS

Anonymous authors
Paper under double-blind review

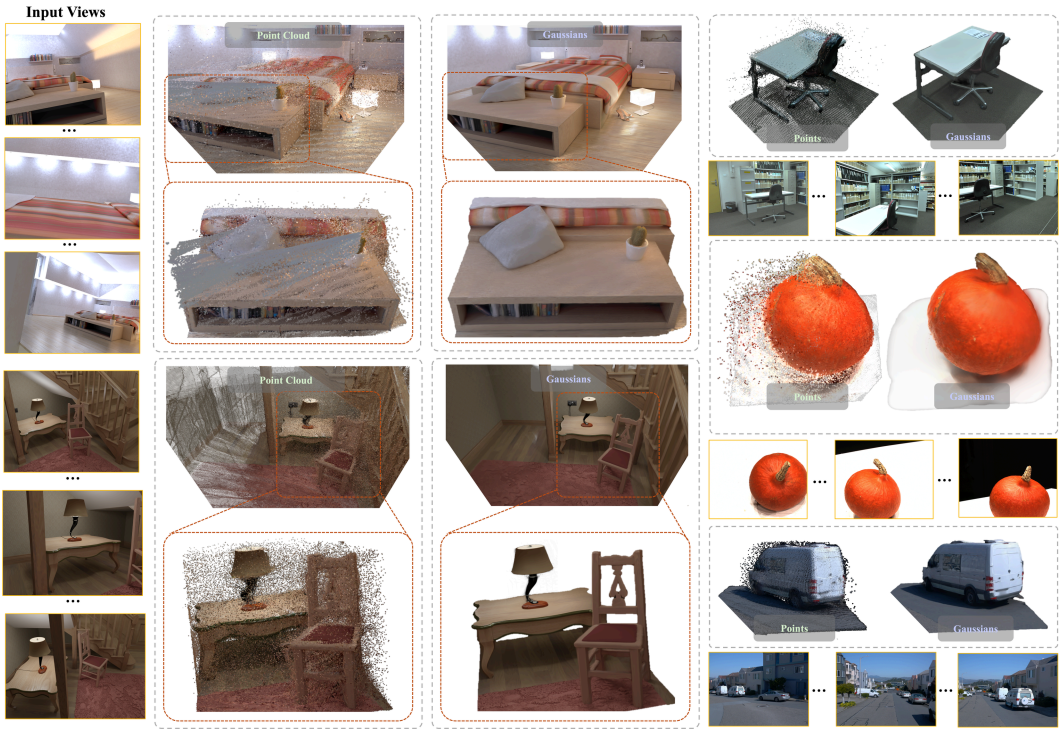

Figure 1: From point clouds to Gaussian scenes. InfiniSplat v2 retains point positions from a geometry foundation model and learns Gaussian attributes for high-quality novel view rendering. Left: input views, point clouds, and corresponding Gaussian renderings, with enlarged regions outlined in orange. Right: additional examples comparing point clouds with the resulting Gaussian scenes.

ABSTRACT

Feed-forward Gaussian reconstruction builds 3D scenes directly from input images for novel view rendering, but learning Gaussian attributes from predicted geometry for high-quality rendering under large viewpoint changes remains challenging. We analyze limitations in Gaussian attribute learning and introduce InfiniSplat v2, which improves rendering quality through architectural and training changes. Our method decouples geometry and attribute prediction with a frozen geometry branch, aligns predicted Gaussians using ground-truth camera poses during training, and adopts large-angle target view supervision. Across six datasets, the feed-forward outputs of the same trained model achieve state-of-the-art mean LPIPS and DISTS among the evaluated methods in both pose-conditioned and pose-free settings. In the pose-conditioned setting, mean LPIPS and DISTS decrease by 25.8% and 32.9%, respectively, compared with the strongest evaluated baselines. We further find that opacity learned with only image reconstruction supervision suppresses geometrically inconsistent points.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Image-based 3D scene reconstruction aims to recover a renderable 3D scene from input images to support novel view rendering for applications such as simulation, content creation, and autonomous driving. 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) enables efficient rendering through an explicit scene representation. Building on this representation, recent feed-forward methods (Charatan et al., 2024; Chen et al., 2024a; Zhang et al., 2024; Xu et al., 2025b) directly predict Gaussian scenes from input images, eliminating the need for per-scene representation optimization at inference time.

Pretrained multi-view geometry models (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a; Yang et al., 2025; Lin et al., 2025; Wang et al., 2026d) provide a useful starting point for constructing renderable scenes from images. They predict depth maps and cameras for each input view. Back-projecting these depth maps with the corresponding camera estimates into a common coordinate frame yields a predicted point cloud. This point cloud captures coherent scene geometry overall, but typically also contains spatially isolated clusters or clusters that are inconsistent across views. Such issues become particularly apparent after 3D aggregation, as illustrated in Fig. 1.

Existing feed-forward methods (Lin et al., 2025; Jiang et al., 2025; Liu et al., 2025; Team et al., 2026) combine geometry prediction from multiple views with Gaussian attribute prediction to construct renderable scenes from input images. For example, DA3-GS (Lin et al., 2025) learns Gaussian attributes on top of the geometry predicted by DA3. However, as shown in Figs. 3 and 4, renderings from DA3-GS and the other evaluated methods still exhibit blur, ghosting, and missing detail. Our work focuses on using these imperfect geometry predictions to construct Gaussian scenes that support high-quality novel view rendering.

As shown in Fig. 1, InfiniSplat v2 reconstructs Gaussian scenes from predicted multi-view geometry and supports high-quality novel-view rendering at extreme viewpoints. We identify key factors that affect Gaussian learning and introduce three corresponding architectural and training improvements. First, we investigate why prior methods struggle to eliminate flying-point artifacts and find that jointly optimizing geometry and Gaussian attributes leads to geometric degradation. We therefore redesign the architecture to decouple geometry prediction from Gaussian attribute prediction while keeping the geometry branch frozen. Second, we find that using predicted camera poses during training (Lin et al., 2025) introduces pose errors. Under large-angle supervision, these errors can misguide the network during training. We therefore use GT camera poses to align the predicted Gaussians. Third, inspired by InfiniSplat's strong performance under large viewpoint changes (Wang et al., 2026b), we introduce large-angle target view supervision for feed-forward Gaussian reconstruction from arbitrary input views. Our experiments show that these three improvements enhance novel view rendering quality under large viewpoint changes, even when the predicted point clouds contain numerous outliers.

We evaluate the same trained model on six datasets, both with and without camera pose inputs. Our method achieves the lowest mean LPIPS and DISTS across the six datasets among the evaluated baselines in both settings. With camera pose inputs, our feed-forward reconstruction reduces LPIPS and DISTS by 25.8% and 32.9%, respectively, relative to the strongest evaluated baseline for each metric. Additionally, we find that, with only image reconstruction supervision, the trainable attribute branch learns to assign lower opacity to outliers that are geometrically inconsistent across views. These points consequently contribute less to the rendered images, reducing visible artifacts in novel views (Fig. 7).

Our contributions are threefold:

- We introduce InfiniSplat v2, a feed-forward Gaussian reconstruction method for arbitrary input views. It freezes the geometry branch to decouple geometry prediction from Gaussian attribute prediction, aligns predicted Gaussians using GT camera poses during training, and adopts large-angle target view supervision.
- We evaluate the method on six datasets in both pose-conditioned and pose-free settings. The feed-forward outputs of the same trained model achieve the lowest mean LPIPS and DISTS among the evaluated baselines in both settings.
- We find that opacity learned with only image reconstruction supervision selectively suppresses the rendering contributions of geometrically inconsistent points across views.

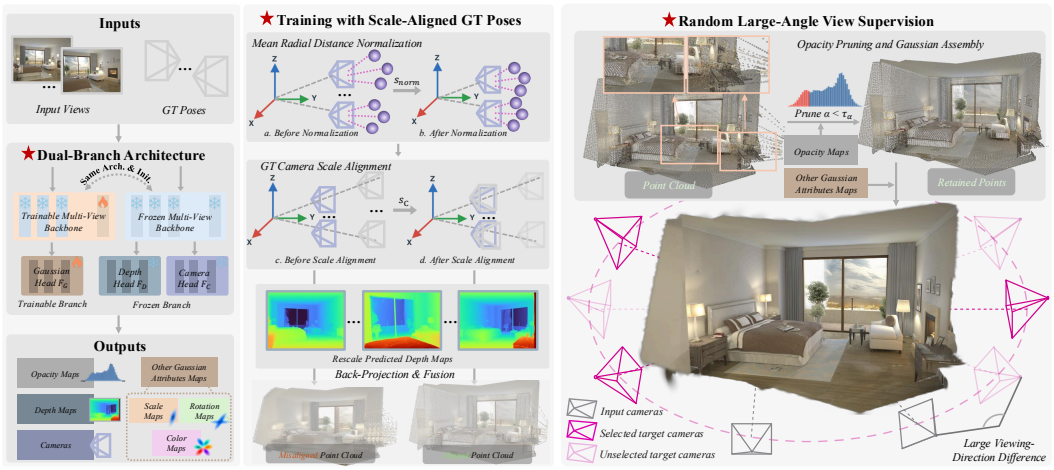

Figure 2: Training pipeline of InfiniSplat v2's feed-forward reconstructor. A frozen geometry branch predicts depth and cameras, while a trainable branch predicts Gaussian attributes. Scale-aligned GT cameras construct the point cloud; after opacity pruning, input and large-angle target views supervise Gaussian learning.

2 Related Work

2.1 Feed-Forward 3D Geometry Estimation

Geometry foundation models recover scene structure from images and provide a geometric basis for feed-forward Gaussian reconstruction. DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), VGGT (Wang et al., 2025a), Fast3R (Yang et al., 2025), and π^3 (Wang et al., 2026d) advance learned geometric reconstruction. Depth Anything 3 (DA3) (Lin et al., 2025) reconstructs geometry from varying numbers of images, with or without supplied poses. VGGT-Long (Deng et al., 2025), VGG-T^3 (Elflein et al., 2026), and Scal3R (Xie et al., 2026) extend learned reconstruction to longer sequences and larger scenes. These predictions provide a useful starting point, but constructing renderable Gaussian scenes also requires learning attributes such as opacity, scale, and appearance.

2.2 Feed-Forward Gaussian Splatting

Generalizable view synthesis learns to predict novel views or renderable representations across scenes (Wang et al., 2021; Hong et al., 2024; Jin et al., 2024). Feed-forward Gaussian methods pursue this goal by directly predicting explicit Gaussian scenes. SHARP (Mescheder et al., 2026) and InfiniSplat (Wang et al., 2026b) reconstruct scenes from a single image, while GS-LRM (Zhang et al., 2024) combines multiple views. NoPoSplat (Ye et al., 2025) and FLARE (Zhang et al., 2025) support reconstruction without supplied camera poses. Long-LRM (Ziwen et al., 2025) and tttLRM (Wang et al., 2026a) further extend reconstruction to long input sequences. Work on Gaussian representations also explores different ways to organize and predict primitives: C3G (An et al., 2025) and TokenGS (Ren et al., 2026) use learnable queries or tokens, while VolSplat (Wang et al., 2025b) predicts voxel-aligned Gaussians.

More directly related to our work is how geometry is estimated and used for Gaussian prediction. pixelSplat (Charatan et al., 2024) establishes cross-view correspondences through epipolar attention, while MVSplat (Chen et al., 2024a) estimates depth using multi-view cost volumes. DepthSplat (Xu et al., 2025b) augments multi-view depth estimation with pretrained monocular depth features. Learned geometric priors also support Gaussian prediction in AnySplat (Jiang et al., 2025), World-Mirror (Liu et al., 2025; Team et al., 2026), and DA3-GS (Lin et al., 2025). Building on such priors, SplatAnything (Anonymous, 2027) scales up training with diverse synthetic and real-world data. QuerySplat (Li et al., 2026) separately decodes geometry and appearance from frozen geometric

features. Our work focuses on learning Gaussian attributes from imperfect predicted point clouds to support high-quality novel view rendering.

After the initial prediction, learned refinement uses rendering feedback to update the Gaussian scene. G3R (Chen et al., 2024b) and the refinement module of Flux4D (Wang et al., 2026c) use explicit rendering gradients to guide learned updates. ReSplat (Xu et al., 2025a) instead uses feature-space rendering errors without explicit gradient feedback. Our refiner uses input-view rendering gradients to learn updates to Gaussian attributes while keeping centers fixed.

3 METHOD

Overview. Given input images and optional camera conditions, InfiniSplat v2 constructs a point cloud from predicted depth maps and corresponding cameras, then learns Gaussian attributes for its points to regulate their rendering contributions. A frozen DA3 branch provides the depth and camera estimates, while a trainable attribute branch adapts to rendering through image reconstruction supervision (Fig. 2). We then refine the scene attributes using input-view rendering feedback.

3.1 BACKGROUND: DA3-BASED GAUSSIAN RECONSTRUCTION

Given N input RGB images $\mathcal{I}_\mathcal{C} = \{\mathbf{I}_i\}_{i\in\mathcal{C}}$ and optional pose conditions $\mathbf{Q}_\mathcal{C}$, our goal is to predict a Gaussian scene for novel view rendering. Here $\mathcal{C}$ indexes the input views. In the pose-free setting, our model receives only RGB images, with no supplied camera intrinsics or poses, and $\mathbf{Q}_\mathcal{C} = \emptyset$. We represent the scene as M Gaussians, separating each center $\boldsymbol{\mu}_m$ from its remaining attributes $\boldsymbol{\xi}_m$:

$$\mathcal{G} = \{(\boldsymbol{\mu}_m, \boldsymbol{\xi}_m)\}_{m=1}^{M}, \qquad \boldsymbol{\xi}_m = (\alpha_m, \mathbf{s}_m, \mathbf{q}_m, \mathbf{c}_m). \tag{1}$$

Here $\boldsymbol{\mu}_m \in \mathbb{R}^3$ specifies position, while $\boldsymbol{\xi}_m$ contains opacity α_m, three-axis scale $\mathbf{s}_m$, rotation $\mathbf{q}_m$, and zeroth-order spherical-harmonic (SH) color $\mathbf{c}_m$. Scale and rotation determine the covariance and thus the Gaussian's shape. For camera $\boldsymbol{\Pi} = (\mathbf{K}, \mathbf{P})$ with intrinsics $\mathbf{K}$ and camera-to-world pose $\mathbf{P}$, the Gaussian renderer (Kerbl et al., 2023) produces $\widehat{\mathbf{I}} = \mathcal{R}(\mathcal{G}; \boldsymbol{\Pi})$.

DA3-GS builds on Depth Anything 3 (DA3) (Lin et al., 2025), using shared multi-view features for depth, camera, and per-pixel Gaussian prediction. Its DPT-style Gaussian head (Ranftl et al., 2021) is trained on a frozen backbone, and an adapter combines predicted geometry and attributes into a renderable scene.

3.2 FEED-FORWARD GAUSSIAN RECONSTRUCTION

Our goal is to improve Gaussian attributes through image reconstruction supervision without compensating for attribute errors by moving point positions. We therefore freeze the geometry branch and keep Gaussian centers fixed at the back-projected point positions. Constructing points with predicted input poses while supervising them from GT target poses can also introduce camera-induced residuals; we use scale-aligned GT cameras for both operations during training. Finally, targets near the input viewing directions may leave Gaussian shape insufficiently constrained. We favor large-angle target views to constrain scales and orientations through projections from different directions. Figure 2 illustrates the resulting training pipeline.

(1) Dual-branch architecture for attribute learning. To preserve DA3's depth and camera estimates while adapting Gaussian attributes for rendering, we use two independently parameterized branches initialized from the same pretrained DA3 weights. Given the same input images and optional pose conditions, the frozen geometry branch predicts depths and cameras, whereas the attribute branch produces multi-layer patch features:

$$\begin{aligned} (\widehat{D}_\mathcal{C}, \widehat{\boldsymbol{\Pi}}_\mathcal{C}) &= \Phi_{\text{geo}}(\mathcal{I}_\mathcal{C}; \mathbf{Q}_\mathcal{C}), \\ \mathbf{F}_\mathcal{C} &= \Phi_{\text{attr}}(\mathcal{I}_\mathcal{C}; \mathbf{Q}_\mathcal{C}). \end{aligned} \tag{2}$$

Only the last K Transformer blocks of Φ_{attr} are fine-tuned; Φ_{geo}, including its depth and camera heads, remains frozen. A DPT Gaussian decoder maps $\mathbf{F}_\mathcal{C}$ to per-pixel attributes $\boldsymbol{\xi}$ (Eq. 1). The points are obtained by the back-projection described below, without predicted positional offsets; scales and rotations remain learnable to adjust Gaussian shape.

(2) Scale-aligned training. Errors in predicted back-projection poses introduce misalignment with GT supervision cameras and add camera-induced residuals to the attribute-learning objective. We therefore use GT cameras for training-time back-projection and align their translation scale with the predicted depths. We first express predicted and GT poses relative to their respective reference cameras, using the same input view as reference. Let $\mathbf{x}_j$ be a point back-projected using predicted depth, intrinsics, and pose, with $\mathcal{V}_{\text{pts}}$ indexing all input pixels. Mean Radial Distance Normalization sets the reconstruction scale:

$$s_{\text{norm}} = \frac{1}{|\mathcal{V}_{\text{pts}}|} \sum_{j \in \mathcal{V}_{\text{pts}}} \|\mathbf{x}_j\|_2, \qquad \bar{D}_i = \frac{\widehat{D}_i}{s_{\text{norm}}}. \tag{3}$$

Both predicted depths and camera translations are divided by s_{norm}, preserving their relative scale. To bring GT cameras to this scale, we fit a similarity transform from GT input-camera centers to the corresponding normalized predicted camera centers using Umeyama (Umeyama, 1991). We retain its scale $s_{\mathcal{C}}$ and preserve GT intrinsics and reference-relative rotations; GT translations become $\widetilde{\mathbf{t}}_v^\star = s_{\mathcal{C}} \mathbf{t}_v^\star$. The same scale is applied to input and target cameras, with only input cameras contributing to the fit. Thus, s_{norm} defines the unit of the predicted reconstruction, while $s_{\mathcal{C}}$ expresses GT camera translations in that unit.

Here stars denote GT quantities and tildes denote scale alignment, with $\widetilde{\mathbf{R}}_i^\star = \mathbf{R}_i^\star$. For each input pixel $\mathbf{u} = (u_x, u_y, 1)^\top$, we back-project the normalized predicted z-depth $\bar{D}_i$ using the scale-aligned GT camera:

$$\boldsymbol{\mu}_{i,\mathbf{u}}^{\text{cand}} = \bar{D}_i(\mathbf{u})\, \widetilde{\mathbf{R}}_i^\star (\mathbf{K}_i^\star)^{-1} \mathbf{u} + \widetilde{\mathbf{t}}_i^\star. \tag{4}$$

The intrinsics $\mathbf{K}_i^\star$ account for image preprocessing. Using scale-aligned GT cameras for both input-view back-projection and target-view rendering preserves their GT relative poses within a common reconstruction frame. This avoids the mismatch caused by constructing points with predicted input poses while supervising them from GT target poses.

We pair these centers with predicted per-pixel attributes and merge redundant points into $\mathcal{G}_{\text{vox}}$ using the voxel-based strategy of AnySplat (Jiang et al., 2025) and WorldMirror 2.0 (Team et al., 2026).

(3) Random large-angle view supervision. Before applying rendering supervision, we prune Gaussians with opacity below τ_α from $\mathcal{G}_{\text{vox}}$, yielding the initial scene $\mathcal{G}^0 = \{(\boldsymbol{\mu}_m^0, \boldsymbol{\xi}_m^0)\}_{m=1}^M$, where M is the retained count. The same merging and pruning are used at inference.

Supervision concentrated near the input viewing directions may leave Gaussian scales and orientations insufficiently constrained. To guide shape learning through projections across views, we favor target cameras whose optical axes differ more from those of the input cameras. For a candidate target $v \notin \mathcal{C}$, let $\mathbf{f}_v = \mathbf{R}_v^\star (0, 0, 1)^\top$ denote its unit optical axis in the common reference frame. We score its angular separation from the nearest input direction by

$$a(v, \mathcal{C}) = \min_{i \in \mathcal{C}} \arccos(\mathbf{f}_v^\top \mathbf{f}_i). \tag{5}$$

We sample the target set $\mathcal{T}$ with a bias toward larger scores, which measure camera orientation differences.

We supervise renderings of $\mathcal{G}^0$ from both input views and the sampled target set $\mathcal{T}$. For target views, GT-depth warp masks restrict the loss to valid regions with depth-consistent support from at least one input view. Target images, cameras, and GT depth serve only sampling and supervision; scene construction and scale estimation use input views alone.

3.3 Learned Rendering-Feedback Refinement

Starting from $\mathcal{G}^0$, a learned refiner updates Gaussian attributes using input-view rendering feedback. Let $\mathcal{G}^t$ denote the scene after t updates. At each step, we render it into all input views, $\mathcal{F} = \mathcal{C}$, and compute

$$\begin{aligned} \mathbf{g}^t &= \text{RenderGrad}(\mathcal{G}^t; \mathcal{I}_{\mathcal{F}}, \boldsymbol{\Pi}_{\mathcal{F}}^{\text{fb}}), \\ \mathbf{h}^t &= E(\mathcal{G}^t, \text{sg}(\mathbf{g}^t), \mathbf{f}). \end{aligned} \tag{6}$$

Here $\boldsymbol{\Pi}_{\mathcal{F}}^{\text{fb}}$ are the input cameras used for feedback. RenderGrad aggregates reconstruction-loss gradients with respect to Gaussian attributes across these views, and sg denotes stop-gradient. The

encoder E combines this feedback with the current scene and projected DPT features $\mathbf{f}$ to form per-Gaussian features $\mathbf{h}^t = \{\mathbf{h}_m^t\}_{m=1}^M$. Projected DPT features are merged using the same voxel groups and weights as the Gaussians to obtain $\mathbf{f}$, which also initializes the recurrent hidden state. The DPT features provide image-derived context and remain fixed across steps, while gradient feedback is recomputed from the current scene. Feedback uses only input observations during both training and inference.

A shared recurrent KNN Transformer U aggregates features across spatial neighbors and predicts attribute residuals and the next hidden state:

$$\begin{aligned}(\Delta\boldsymbol{\xi}_m^t, \mathbf{b}_m^{t+1}) &= U(\boldsymbol{\xi}_m^t, \mathbf{h}_m^t, \mathbf{b}_m^t),\\ \boldsymbol{\xi}_m^{t+1} &= \mathrm{Apply}(\boldsymbol{\xi}_m^t, \Delta\boldsymbol{\xi}_m^t), \qquad \boldsymbol{\mu}_m^{t+1} = \boldsymbol{\mu}_m^0.\end{aligned} \tag{7}$$

Here $\mathbf{b}_m^t$ is the recurrent hidden state, and Apply applies residuals in each attribute's parameterization space. Gaussian centers and counts remain fixed. Rendering gradients serve as inputs to learned updates rather than direct gradient-descent steps. The fixed centers allow the spatial KNN graph to be reused across steps. The final scene is denoted $\mathcal{G}^{\mathrm{ref}}$.

3.4 Training and Inference

Training. In Stage I, both branches receive input RGB and GT pose conditions. We supervise input and sampled target renderings with equally weighted masked L_1 and LPIPS (Zhang et al., 2018) losses, updating only the attribute branch's last K Transformer blocks and Gaussian decoder. Stage II freezes the reconstructor and trains the feature projection and refiner. Input-view losses provide feedback, while input and disjoint target views supervise the updated scene.

Inference. The same trained model supports pose-conditioned and pose-free inference. In the pose-conditioned setting, we construct centers using the supplied intrinsics and poses, rescaling their translations to match the normalized predicted depths. In the pose-free setting, we use the geometry branch's predicted cameras, normalizing their translations by s_{norm}. In either setting, these cameras also render the input views for refinement feedback. The refiner updates scene attributes while network weights remain fixed.

The appendix specifies network architectures, gradient handling, camera conventions, sampling, losses, and training hyperparameters.

4 Experiments

4.1 Experimental Setup

Datasets and baselines. We train InfiniSplat v2 on a mixture of Hypersim (Roberts et al., 2021), Raw3D (Team, 2026), and ScanNet++ (Yeshwanth et al., 2023), using disjoint scenes for training and validation. We evaluate on six datasets: DL3DV (Ling et al., 2024), ETH3D (Schops et al., 2017), NRGBD (Azinović et al., 2022), Mip-NeRF 360 (Barron et al., 2022), RealEstate10K (RE10K) (Zhou et al., 2018), and DTU (Aanæs et al., 2016). We compare with tttLRM (Wang et al., 2026a), AnySplat (Jiang et al., 2025), C3G (An et al., 2025), DA3-GS (Lin et al., 2025), DepthSplat (Xu et al., 2025b), QuerySplat (Li et al., 2026), ReSplat (Xu et al., 2025a), TokenGS (Ren et al., 2026), VolSplat (Wang et al., 2025b), and WorldMirror 2.0 (Team et al., 2026).

Evaluation protocol. We use $N \in \{4, 8, 16\}$ input views and four held-out target views per example, with the same view splits for all methods. We evaluate methods separately in pose-conditioned and pose-free settings. We use the same trained model for both settings. Ours denotes the feed-forward output, and Ours* includes three learned refinement steps. Evaluation-camera alignment is separate from reconstruction and uses only input-camera correspondences. Following SHARP (Mescheder et al., 2026), we use LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020) as our primary metrics for perceptual quality. Within each dataset, scores are computed on the same target views and valid regions. Per-dataset scores average equally over the three input-view counts, and the overall mean weights the six datasets equally. The complete camera and scoring protocol is provided in the appendix.

Table 1: Novel view synthesis on six datasets. Scores average $N = 4, 8, 16$ equally; Mean weights datasets equally. Ours/Ours*: feed-forward/refined. First, second, and third are highlighted per input setting; best in bold.

Input	Method	DL3DV		ETH3D		NRGBD		Mip-NeRF 360		RE10K		DTU		Mean	
		LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓
Pose-free	AnySplat	0.5985	0.2490	0.5473	0.2297	0.1895	0.1131	0.5512	0.2259	0.2454	0.1113	0.4800	0.2345	0.4353	0.1939
	C3G	0.7503	0.3694	0.6722	0.3471	0.5393	0.2708	0.7390	0.3878	0.4600	0.2145	0.7115	0.3704	0.6454	0.3267
	QuerySplat	0.5601	0.2260	0.5805	0.2750	0.2586	0.1405	0.6076	0.2678	0.2158	0.1113	0.5199	0.2722	0.4571	0.2155
	DA3-GS	0.5146	0.2179	0.4950	0.2435	0.3456	0.1777	0.4738	0.2303	0.2526	0.1434	0.4108	0.2336	0.4154	0.2077
	WorldMirror 2.0	0.4800	0.2148	0.4917	0.2310	0.2099	0.1556	0.4701	0.2169	0.2210	0.1323	**0.3812**	0.2106	0.3757	0.1935
	Ours	0.3773	**0.1293**	0.4145	**0.1710**	0.1096	0.0788	**0.3784**	**0.1504**	0.1459	0.0819	0.3898	**0.1932**	0.3026	**0.1341**
	Ours*	**0.3698**	0.1306	**0.4020**	0.1738	**0.1033**	**0.0776**	0.3788	0.1595	**0.1445**	**0.0808**	0.3835	0.2048	**0.2970**	0.1379
Pose-conditioned	tttLRM	0.4955	0.2574	0.4984	0.2296	0.2857	0.1979	0.3835	0.2036	0.2193	0.1404	0.3933	0.2662	0.3793	0.2159
	DA3-GS	0.5235	0.2138	0.5776	0.2443	0.3593	0.1557	0.4965	0.2390	0.2315	0.1263	0.3536	0.2000	0.4237	0.1965
	DepthSplat	0.5766	0.2707	0.5615	0.2874	0.0867	0.0673	0.4504	0.2448	0.1297	0.0840	0.4263	0.2413	0.3718	0.1993
	ReSplat	0.4198	0.2123	0.5266	0.2309	0.1204	0.0962	0.3655	0.1747	0.1191	0.0842	0.4485	0.2781	0.3333	0.1794
	TokenGS	0.6911	0.3239	0.6704	0.2913	0.5589	0.3025	0.6456	0.3095	0.3955	0.2094	0.5057	0.3221	0.5779	0.2931
	VolSplat	0.7049	0.3539	0.6715	0.3149	0.2342	0.1911	0.5071	0.2585	0.3552	0.1985	0.6424	0.3370	0.5192	0.2756
	WorldMirror 2.0	0.4673	0.2126	0.5201	0.2079	0.1056	0.0717	0.4689	0.2194	0.2174	0.1359	0.3306	0.1930	0.3517	0.1734
	Ours	0.3363	0.1209	0.4339	0.1661	0.0653	0.0535	0.3206	0.1405	0.1202	**0.0798**	0.2070	**0.1368**	0.2472	0.1163
	Ours*	**0.3276**	**0.1197**	**0.4176**	**0.1631**	**0.0612**	**0.0524**	**0.3099**	**0.1380**	**0.1176**	0.0802	**0.2014**	0.1422	**0.2392**	**0.1159**

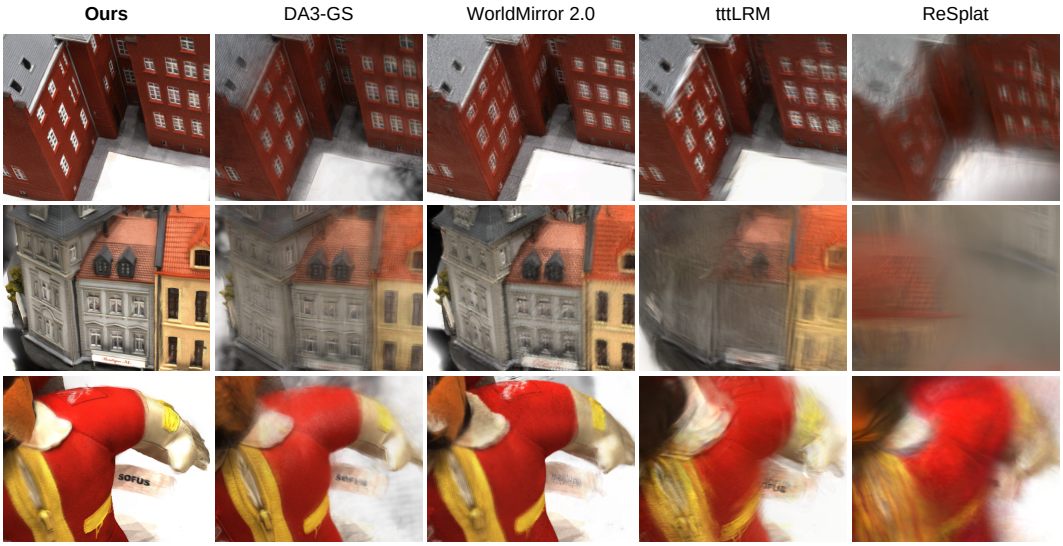

Figure 3: Object-level qualitative comparisons under multi-view input settings.

4.2 Comparison with State-of-the-Art Methods

Quantitative comparison. Across six datasets, our feed-forward model achieves the best mean LPIPS and DISTS among evaluated baselines in both settings (Table 1). With pose conditioning, LPIPS improves by 25.8% over ReSplat and DISTS by 32.9% over WorldMirror 2.0; without poses, the gains over WorldMirror 2.0 are 19.5% and 30.7%, respectively. Results for all four metrics and a comparison of learned refinement with Adam are provided in the appendix.

Qualitative evaluation. Figures 3 and 4 show that our method preserves fine textures and structural details in both object-level and scene-level reconstructions. More qualitative results are provided in the appendix.

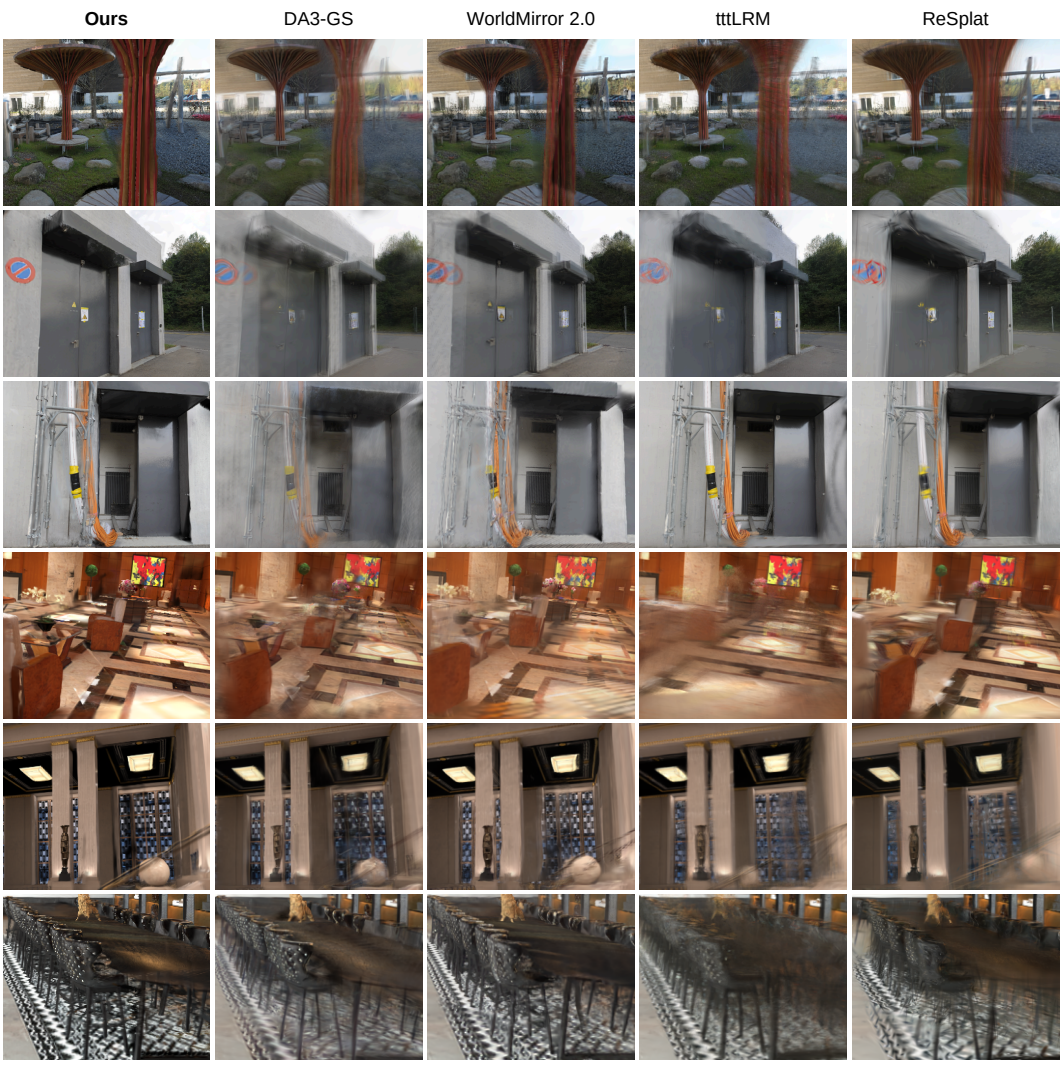

Figure 4: Scene-level qualitative comparisons under multi-view input settings.

4.3 Ablations and Analysis

We evaluate attribute-backbone fine-tuning (A), geometry freezing (G), GT-camera backprojection (C), and large-angle supervision (V) on ETH3D and DTU, averaging pre-refinement results over $N = 4, 8, 16$ input views. We additionally evaluate learned positional offsets (w/ O). Ours uses $K = 10$; w/o A uses $K = 0$.

Under image reconstruction supervision, geometry fine-tuning (w/o G) and learned offsets (w/ O) worsen both metrics on ETH3D and DTU (Table 2). Freezing geometry instead of fine-tuning reduces LPIPS by 16.6% and 23.2%, respectively. We interpret this degradation as positional updates compensating for attribute errors, a pathway blocked by fixed centers. With the geometry backbone frozen, w/ O predicts 3D offsets to back-projected centers through the Gaussian head, producing local blur and depth distortions in the NRGBD kitchen example (Fig. 5).

Fine-tuning the last $K = 10$ attribute-backbone blocks, compared with $K = 0$ (w/o A), reduces LPIPS by 1.1% on ETH3D and 20.8% on DTU, with lower DISTS on both datasets. GT-camera backprojection reduces DTU LPIPS by 36.1% relative to predicted cameras (w/o C). Large-angle target view supervision improves both LPIPS and DISTS on ETH3D and DTU compared with local interpolation-view supervision (w/o V).

Variant	ETH3D		DTU	
	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓
Ours	**0.4339**	**0.1661**	**0.2070**	**0.1368**
w/o A	0.4388	0.2019	0.2612	0.1663
w/o G	0.5204	0.1938	0.2695	0.1841
w/ O	0.4998	0.2082	0.2826	0.1887
w/o C	0.5017	0.1863	0.3237	0.1758
w/o V	0.4533	0.1763	0.2295	0.1508

Table 2: Ablation study of reconstruction components on ETH3D and DTU.

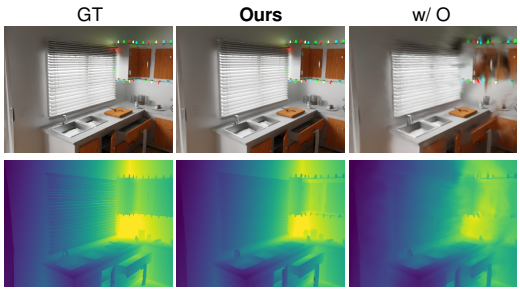

Figure 5: Effect of positional offsets on rendering and depth in an NRGBD scene.

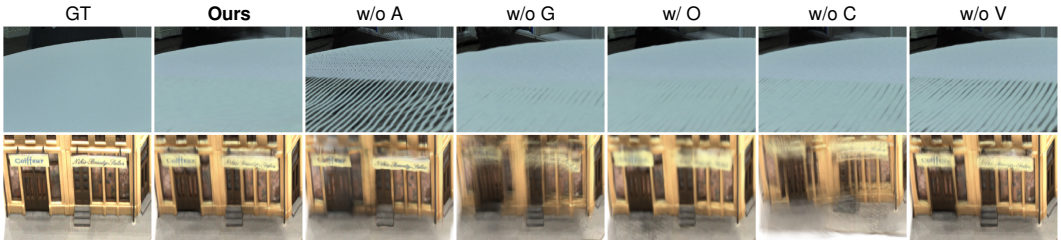

Figure 6: Effect of reconstruction components on surface artifacts and structural detail.

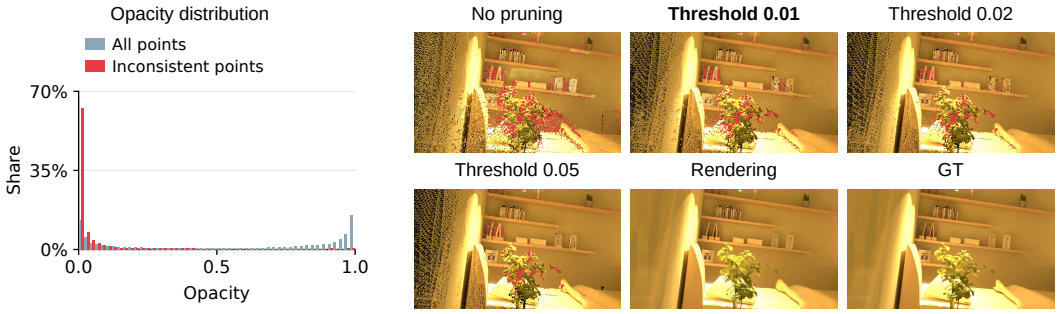

Figure 7: Opacity distributions and the effect of pruning on geometrically inconsistent points. GT depth independently defines the inconsistent subset; distributions are normalized separately.

Qualitative ablations. In Fig. 6, w/o A and w/o V produce pronounced cracks on the tabletop in the upper row; w/o G, w/ O, and w/o C exhibit milder cracks. In the lower row, w/o G and w/o C blur the door frames, which remain more distinct in the full model.

Opacity and geometric consistency. Using GT-depth labels on held-out Hypersim (Roberts et al., 2021) validation scenes, we analyze opacity learned with only image reconstruction supervision. Labels are independent of opacity: inconsistent points violate GT-observed free space in at least two views and agree with no observed surface depth. At threshold 0.01, pruning removes 51.8% of geometrically inconsistent points but only 7.5% of all points, retaining 93.3% of surface-consistent points. This selective removal links opacity to geometric reliability while preserving the main scene structure and appearance (Fig. 7).

5 Conclusion

InfiniSplat v2 improves mean LPIPS and DISTS across six datasets in both pose settings. ETH3D and DTU ablations support fixed centers and attribute adaptation under reconstruction supervision. Learned opacity preferentially suppresses geometrically inconsistent points.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing of author-written text and with code development and debugging. The authors reviewed and approved the revised text and inspected and tested the AI-assisted code. Generative AI tools were not used to originate the research ideas or core technical contributions, or to generate experimental or training data. References were independently identified, collected, and selected by the authors. The authors take full responsibility for the final content of this work, including its text, claims, experimental results, and code.

REFERENCES

Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2): 153–168, 2016.

Honggyu An, Jaewoo Jung, Mungyeom Kim, Sunghwan Hong, Chaehyun Kim, Kazumi Fukuda, Minkyeong Jeon, Jisang Han, Takuya Narihira, Hyuna Ko, et al. C3g: Learning compact 3d representations with 2k gaussians. *arXiv preprint arXiv:2512.04021*, 2025.

Anonymous. Splatanything: Gaussian spacing normalization for feed-forward 3d gaussian splatting across arbitrary scenes, 2027. Manuscript under review at ICLR 2027.

Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6290–6301, June 2022.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5460–5469. IEEE, 2022.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19457–19467. IEEE, 2024.

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*, pp. 370–386. Springer, 2024a.

Yun Chen, Jingkang Wang, Ze Yang, Sivabalan Manivasagam, and Raquel Urtasun. G3r: Gradient guided generalizable reconstruction. In *European Conference on Computer Vision*, pp. 305–323. Springer, 2024b.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. *arXiv preprint arXiv:2507.16443*, 2025.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. VGG-T3: Offline feed-forward 3d reconstruction at scale. *arXiv preprint arXiv:2602.23361*, 2026.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2485–2494, 2020.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations*, volume 2024, pp. 50678–50702, 2024.

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. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024.

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

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Yinglong Li, Donghui Shen, Xiaoyu Zhang, Zhichao Ye, Hongyu Wu, Aimin Hao, Guofeng Zhang, and Haomin Liu. Querysplat: Decoupling geometry and appearance representations in 3dgs prediction. *arXiv preprint arXiv:2608.01186*, 2026.

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

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025.

Lars Mescheder, Wei Dong, Shiwei Li, Xuyang Bai, Marcel Santos, Peiyun Hu, Bruno Lecouat, Mingmin Zhen, Amaël Delaunoy, Tian Fang, et al. Sharp monocular view synthesis in less than a second. In *International Conference on Learning Representations*, volume 2026, pp. 24192–24230, 2026.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Jiawei Ren, Michal Jan Tyszkiewicz, Jiahui Huang, and Zan Gojcic. Tokengs: Decoupling 3d gaussian prediction from pixels with learnable tokens. *arXiv preprint arXiv:2604.15239*, 2026.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10892–10902. IEEE, 2021.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451. IEEE, 2020.

HY Team, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, et al. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv e-prints*, pp. arXiv–2604, 2026.

Rawmantic Team. Rawmantic 3d training field: From the real world, for the real world. *Rawmantic*, 2026.

Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12912–12921. IEEE, 2022.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 1991.

Chen Wang, Hao Tan, Wang Yifan, Zhiqin Chen, Yuheng Liu, Kalyan Sunkavalli, Sai Bi, Lingjie Liu, and Yiwei Hu. tttlrm: Test-time training for long context and autoregressive 3d reconstruction. *arXiv preprint arXiv:2602.20160*, 2026a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Jiawei Wang, Hao Yu, Yongzhen Hu, Xinyi Yang, Tao Ni, Xin Zhan, Junbo Chen, Xiaowei Zhou, Ruizhen Hu, and Sida Peng. Infinisplat: Implicit gaussian decoding for large-baseline monocular view synthesis. *arXiv preprint arXiv:2608.02437*, 2026b.

Jingkang Wang, Henry Che, Yun Chen, Ze Yang, Lily Goli, Sivabalan Manivasagam, and Raquel Urtasun. Flux4d: Flow-based unsupervised 4d reconstruction. *Advances in Neural Information Processing Systems*, 38:105673–105693, 2026c.

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2021.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Weijie Wang, Yeqing Chen, Zeyu Zhang, Hengyu Liu, Haoxiao Wang, Zhiyuan Feng, Wenkang Qin, Feng Chen, Jia-Wang Bian, Zheng Zhu, et al. Volsplat: Rethinking feed-forward 3d gaussian splatting with voxel-aligned prediction. *arXiv preprint arXiv:2509.19297*, 2025b.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Pi3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026d.

Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, et al. Scal3r: Scalable test-time training for large-scale 3d reconstruction. *arXiv preprint arXiv:2604.08542*, 2026.

Haofei Xu, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Resplat: Learning recurrent gaussian splatting. *arXiv preprint arXiv:2510.08575*, 2025a.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16453–16463. IEEE, 2025b.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935. IEEE, 2025.

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. In *International Conference on Learning Representations*, volume 2025, pp. 54009–54033, 2025.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22. IEEE, 2023.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21936–21947. IEEE, 2025.

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.

Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4349–4359. IEEE, 2025.

A Detailed Experimental Setup

A.1 Data and View Sampling

Training data and inputs. Both stages sample Hypersim (Roberts et al., 2021), Raw3D (Team, 2026), and ScanNet++ (Yeshwanth et al., 2023) with probabilities of 60%, 10%, and 30%, using disjoint training and validation scenes. Each sample contains four target views and $N \in \{4, \ldots, 12\}$ input views. The processed image's longer side is 1022 pixels for $N \leq 8$, 756 for $N = 9, 10$, and 518 for $N = 11, 12$.

Target-view sampling. We rank non-input views by decreasing minimum optical-axis angle $a(v, \mathcal{C})$ (Eq. 5), assigning the upper and lower halves to large- and small-angle pools. Each sample selects the large-angle pool with probability 0.7 or the small-angle pool with probability 0.3, then draws four target views uniformly without replacement. The selected pool must contain at least four candidates. Figure 8 shows the sampled angular distributions.

Training masks. Target supervision uses visibility masks constructed from GT depth. A target pixel must have finite GT depth greater than 10^{-6} and pass the supplied validity mask. After back-projection, it is retained if at least one input view places it in front of the camera and within the image bounds, has valid sampled depth, and satisfies

$$|z_{\mathrm{proj}} - d_{\mathrm{src}}| \leq \max(\tau_{\mathrm{abs}}, \tau_{\mathrm{rel}} d_{\mathrm{src}}). \tag{8}$$

Here z_{proj} is the point depth in that input camera and d_{src} is the sampled depth. The mask function defaults to $\tau_{\mathrm{abs}} = 0.03$ and $\tau_{\mathrm{rel}} = 0.01$.

Local-view control. The w/o V variant samples local interpolation views between input views, retaining target supervision and the full model's mask construction.

Evaluation views. We evaluate on six datasets with $N = 4, 8, 16$ input views.

A.2 Architecture, Rendering, and Coordinate Conventions

Initialization and rendering. Both branches are initialized from DA3NESTED-GIANT-LARGE-1.1 with independent parameters. Feature projection and a DPT decoder predict Gaussian attributes. Voxel merging uses a side length of 0.004 in the scale-normalized reconstruction frame. During training and inference, Gaussians with opacity below $\tau_\alpha = 0.01$ are pruned after merging and before refinement. All experiments use gsplat with `eps2d`=10^{-2}.

Scale alignment and point construction. Camera alignment applies only a scale factor, not the full similarity transform. Predicted and GT poses are expressed relative to their respective reference cameras, using the first input view in sorted order. Predicted depths and translations are normalized by $s_{\mathrm{norm}} > 0$ from Eq. 3. Using only input-camera center correspondences, we fit an Umeyama similarity transform (Umeyama, 1991) from the GT trajectory to the normalized predicted trajectory and retain its scale $s_{\mathcal{C}}$. This scale applies to both input and target translations, preserving GT intrinsics and reference-relative rotations:

$$\widetilde{\mathbf{R}}_v^\star = \mathbf{R}_v^\star, \qquad \widetilde{\mathbf{t}}_v^\star = s_{\mathcal{C}} \mathbf{t}_v^\star. \tag{9}$$

During training, Gaussian centers are constructed from normalized predicted z-depth, GT intrinsics, and scale-aligned GT poses.

A.3 Optimization Settings

Both stages use GT pose conditions and AdamW on 16 H20 GPUs, with $\beta = (0.9, 0.95)$, a global batch size of 16, and global L_2 gradient clipping at 1.0. Each GPU processes one multi-view sample without gradient accumulation. Stage I trains only the last ten attribute-backbone Transformer blocks and the Gaussian head, with peak learning rates of 10^{-6} and 2×10^{-4}, respectively. Training runs for 25,000 steps, with 2,000 steps of linear warmup followed by cosine decay to 0.1 times the peaks, taking about 48 hours. Stage II freezes the reconstructor and trains the DPT feature projection and refiner for 5,000 steps. The learning rate starts at 2×10^{-4} without warmup and decays by a cosine schedule to 0.1 times its initial value, taking about 13 hours.

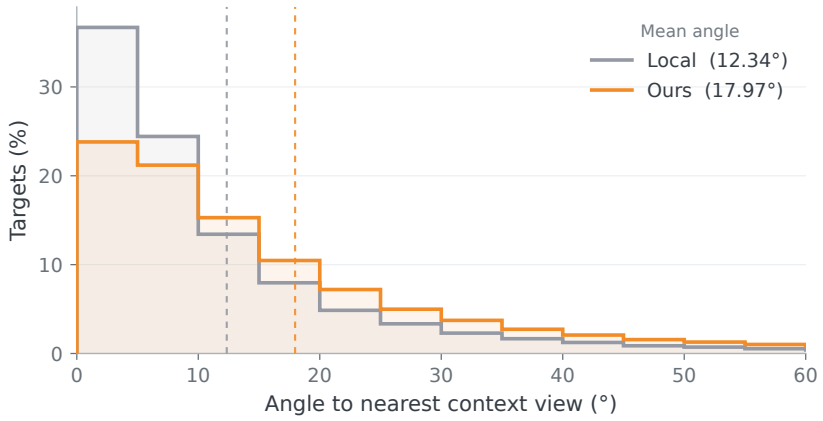

Figure 8: Target-view angle distributions. Dashed lines mark the mean minimum angle to input views: 12.34° for local sampling and 17.97° for ours.

A.4 Stage-I Reconstruction Loss

We render the initial scene $\mathcal{G}^0$ into input and target views $v \in \mathcal{V} = \mathcal{C} \cup \mathcal{T}$:

$$\widehat{\mathbf{I}}_v = \mathcal{R}(\mathcal{G}^0; \widetilde{\mathbf{\Pi}}^\star_v). \tag{10}$$

Let M_v be the valid supervision mask and $e_v(p)$ the RGB-channel-averaged absolute error at pixel p. LPIPS (Zhang et al., 2018) uses VGG features, with invalid regions zeroed before input normalization, yielding the spatial distance map

$$d_v = \mathrm{LPIPS}_{\mathrm{spatial}}(M_v \odot \widehat{\mathbf{I}}_v,\ M_v \odot \mathbf{I}_v). \tag{11}$$

The two losses receive equal weights and are aggregated over valid pixels:

$$\mathcal{L}_{\mathrm{I}} = \frac{\sum_{v\in\mathcal{V}} \sum_p M_v(p)[e_v(p) + d_v(p)]}{\max(1, \sum_{v\in\mathcal{V}} \sum_p M_v(p))}. \tag{12}$$

A.5 Stage-II Learned Refinement

Features and architecture. DPT features are projected to eight channels by a learnable 1×1 convolution, merged using the Gaussian voxel groups and weights, and reused across steps. A linear projection initializes the 16-dimensional hidden state. Image features, current attributes, gradient feedback, neighbor distances, and hidden states are concatenated and projected to 64 dimensions, then processed by four KNN Transformer blocks. Each block contains single-head attention over eight neighbors, LayerNorm, an MLP, and residual connections. The output is concatenated with the original input, and an MLP predicts attribute and hidden-state increments.

Feedback and attribute updates. During training and inference, all input views provide feedback. At each step, the current scene is rendered into these views with feedback loss

$$\ell_{\mathrm{fb}} = 0.8L_1 + 0.2L_{\mathrm{DSSIM}}, \qquad L_{\mathrm{DSSIM}} = (1 - \mathrm{SSIM})/2. \tag{13}$$

DSSIM is computed over valid image regions. Gradients with respect to Gaussian attributes are normalized and detached before entering the shared updater. Gaussian centers, counts, and correspondence remain fixed. Scales update as $s_{\mathrm{new}} = s_{\mathrm{old}} \exp(\Delta \log s)$ without residual clipping. Opacity-logit, local rotation-vector, and RGB-logit increments use elementwise $a \tanh(\mathrm{raw})$, with $a = 0.1$, 0.25, and 1.0, respectively; hidden-state increments use $0.1 \tanh(\mathrm{raw})$. Logit inputs are clamped to $[10^{-6}, 1 - 10^{-6}]$, with no additional cross-step residual decay. Both stages use zeroth-order SH colors; the refiner updates colors in RGB-logit coordinates before converting them back to SH coefficients.

Training unroll and supervision. Stage II uses the same input-view counts and resolutions as Stage I, sampling K_{r} uniformly from $\{1, 2, 3\}$ updates during training and using three updates at inference. Input and target views supervise updated scenes. Feedback gradients and cross-step

Gaussian attribute paths are detached, while hidden-state residual updates remain differentiable for recurrent backpropagation. The outer loss for each supervised view is

$$\ell_{\text{out}} = 0.8L_1 + 0.2L_{\text{DSSIM}} + 0.5L_{\text{LPIPS}}. \tag{14}$$

Input and target groups each receive weight 0.5, with equal weights within each group; invalid views are skipped and remaining weights renormalized. Let $\mathcal{L}^k_{\text{render}}$ denote the grouped outer loss after update k. The multi-step objective is

$$\mathcal{L}_{\text{II}} = \sum_{k=1}^{K_{\text{r}}} 0.9^{K_{\text{r}}-k}\mathcal{L}^k_{\text{render}} + 0.1\mathcal{L}_{\text{opacity}}. \tag{15}$$

Step weights are not normalized. Opacity regularization penalizes the change in logits between final and initial Gaussians:

$$\mathcal{L}_{\text{opacity}} = \frac{1}{M}\sum_{m=1}^{M}[\text{logit}_\epsilon(\alpha_m^{K_{\text{r}}}) - \text{logit}_\epsilon(\alpha_m^0)]^2, \qquad \epsilon = 10^{-6}. \tag{16}$$

First-order surrogate loss. Let x^k collect updated log-scale, opacity-logit, local-rotation, and RGB-logit parameters. Rendering gradients are computed on independent leaf copies $z^k = \text{sg}(x^k)$ to form

$$\begin{aligned} r^k &= \text{sg}(\nabla_z \mathcal{L}_{\text{render}}(z^k)), \\ \mathcal{L}^k_{\text{proxy}} &= \text{sg}(\mathcal{L}_{\text{render}}(z^k)) + \langle x^k - \text{sg}(x^k), r^k\rangle. \end{aligned} \tag{17}$$

Here sg preserves values while stopping gradients. The surrogate equals the rendering loss in value, with gradient

$$\nabla_\theta \mathcal{L}^k_{\text{proxy}} = \left(\frac{\partial x^k}{\partial \theta}\right)^\top r^k \tag{18}$$

for trainable parameters θ. Per-step surrogates follow the multi-step objective weights without second-order differentiation.

Chunked execution. The global KNN graph is reused across steps. After stable sorting along the longest voxel-coordinate axis, core sets of at most 200,000 points are expanded with four-hop neighborhoods. Chunks exceeding 300,000 points are recursively bisected by point count along their core sets' longest axes, with neighborhoods recomputed. Only core-point outputs are retained; training uses gradient checkpointing.

A.6 Camera Protocol and Evaluation Metrics

Camera inputs and rendering. Pose-conditioned and pose-free results are reported and ranked separately. In the pose-free setting, reference input cameras are not reconstruction inputs; they only establish the evaluation reference frame and estimate translation scale from input-camera correspondences. The scene is constructed before evaluation-camera alignment and remains unchanged. Target poses define rendering queries without entering the fit. Alignment rescales reference-relative translations only, rather than applying a full Sim(3) transform. For methods predicting intrinsics, target focal lengths are the component-wise median of predicted input focal lengths, scaled to the target resolution, with the reference principal point retained. In the pose-conditioned setting, reference input cameras are available for reconstruction and target intrinsics are fixed to GT. Evaluation target RGB is never used for reconstruction or refinement feedback.

Metrics and aggregation. We evaluate novel view rendering with PSNR, SSIM, LPIPS, and DISTS. Both settings average equally at each successive level: target views within each example, examples within each scene, scenes, the three input configurations, and the six datasets.

B Additional Experiments and Results

B.1 Full Quantitative Results

Six-dataset means. Tables 3 and 4 summarize PSNR, SSIM, LPIPS, and DISTS on DL3DV, ETH3D, NRGBD, Mip-NeRF 360, RE10K, and DTU. Our refined output ranks first on all four mean metrics in the pose-conditioned setting. In the pose-free setting, refinement gives the best mean PSNR, SSIM, and LPIPS, while the feed-forward output has the lowest mean DISTS.

Table 3: Pose-free results averaged over six datasets.

Method	PSNR↑ (dB)	SSIM↑	LPIPS↓	DISTS↓
AnySplat	15.6668	0.5531	0.4353	0.1939
C3G	13.2165	0.5653	0.6454	0.3267
QuerySplat	17.0532	0.6404	0.4571	0.2155
DA3-GS	17.7144	0.6463	0.4154	0.2077
WorldMirror 2.0	17.0031	0.6089	0.3757	0.1935
Ours	18.4002	0.6423	0.3026	**0.1341**
Ours*	**18.8483**	**0.6523**	**0.2970**	0.1379

Table 4: Pose-conditioned results averaged over six datasets.

Method	PSNR↑ (dB)	SSIM↑	LPIPS↓	DISTS↓
tttLRM	17.8632	0.6518	0.3793	0.2159
DA3-GS	16.8535	0.6100	0.4237	0.1965
DepthSplat	18.9085	0.6602	0.3718	0.1993
ReSplat	20.9139	0.7117	0.3333	0.1794
TokenGS	15.9149	0.5990	0.5779	0.2931
VolSplat	14.2476	0.5060	0.5192	0.2756
WorldMirror 2.0	17.7222	0.6192	0.3517	0.1734
Ours	20.2824	0.6988	0.2472	0.1163
Ours*	**20.9334**	**0.7131**	**0.2392**	**0.1159**

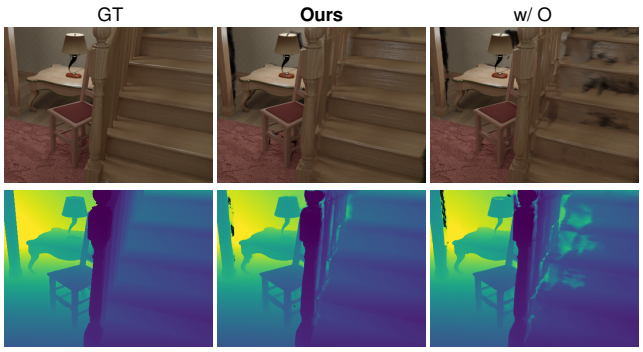

Figure 9: Positional offsets on NRGBD with eight input views. Top: RGB; bottom: depth. Rendered depth is alpha-normalized expected camera-z depth, without fitting to target GT. Colors indicate near (purple), far (yellow), and invalid (gray) pixels.

Results by input-view count. Tables 5, 6, and 7 report per-dataset LPIPS and DISTS with four, eight, and sixteen input views. All tables follow the colors of Table 1, ranking unrounded scores within each pose setting and bolding the best; dataset means are equally weighted.

B.2 Feed-Forward Reconstructor Ablations

All variants are evaluated on ETH3D and DTU before refinement, averaging equally over $N = 4, 8, 16$, with other training settings unchanged. The w/o A variant freezes the attribute backbone ($K = 0$), shares backbone features across branches, and trains only the Gaussian head. The w/o G variant fine-tunes the geometry backbone at 10^{-6}, keeping the Gaussian head learning rate at 2×10^{-4}. The w/ O variant keeps the geometry backbone frozen and predicts 3D offsets added to back-projected centers in the scale-aligned frame. The w/o C variant uses predicted cameras for training-time back-projection. The w/o V variant uses local-view sampling, retaining target supervision and the same training masks. Figure 9 provides an NRGBD staircase comparison, where the offset variant shows stronger surface artifacts and depth deviations than the fixed-center model.

B.3 Learned Refinement versus Direct Gradient Descent

We compare feed-forward outputs, Adam, and learned refinement on ETH3D, averaging equally over $N = 4, 8, 16$. Both refinements start from the same scene and perform three updates to scales,

Table 5: Per-dataset results with $N = 4$ input views. Mean weights datasets equally.

Input	Method	DL3DV		ETH3D		NRGBD		Mip-NeRF 360		RE10K		DTU		Mean	
		LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓
Pose-free	AnySplat	0.6004	0.2527	0.5244	0.2245	0.1917	0.1153	0.5259	0.2187	0.2386	0.1084	0.4995	0.2452	0.4301	0.1941
	C3G	0.7565	0.3785	0.6090	0.2877	0.6137	0.3191	0.6849	0.3448	0.5080	0.2671	0.7031	0.3680	0.6459	0.3275
	QuerySplat	0.5453	0.2191	0.5315	0.2425	0.2580	0.1410	0.5621	0.2429	0.2172	0.1117	0.5243	0.2759	0.4397	0.2055
	DA3-GS	0.5347	0.2145	0.4937	0.2291	0.3352	0.1703	0.4933	0.2199	0.2412	0.1393	0.4287	0.2411	0.4211	0.2024
	WorldMirror 2.0	0.4647	0.2021	0.4484	0.2107	0.1793	0.1386	0.4432	0.1957	0.1903	0.1166	**0.3924**	0.2076	0.3530	0.1786
	Ours	0.3862	0.1322	0.3850	**0.1596**	0.1078	0.0779	0.3583	**0.1390**	0.1438	0.0806	0.4162	**0.1938**	0.2995	**0.1305**
	Ours*	**0.3774**	**0.1309**	**0.3718**	0.1596	**0.1009**	**0.0765**	**0.3547**	0.1422	**0.1416**	**0.0800**	0.4176	0.2029	**0.2940**	0.1320
Pose-conditioned	tttLRM	0.5957	0.2962	0.5477	0.2476	0.3496	0.2262	0.4451	0.2349	0.2554	0.1541	0.4803	0.3105	0.4456	0.2449
	DA3-GS	0.5265	0.2098	0.6205	0.2595	0.3678	0.1548	0.4851	0.2200	0.2320	0.1248	0.3731	0.2083	0.4342	0.1962
	DepthSplat	0.5986	0.2723	0.5085	0.2316	0.0639	0.0550	0.3872	0.1952	**0.1066**	**0.0734**	0.4907	0.2622	0.3592	0.1816
	ReSplat	0.4309	0.2062	0.5511	0.2439	0.0708	0.0755	0.3676	0.1760	0.1073	0.0788	0.4884	0.2923	0.3360	0.1788
	TokenGS	0.7631	0.3685	0.6764	0.2972	0.5966	0.3103	0.6393	0.3097	0.4131	0.2165	0.5752	0.3505	0.6106	0.3088
	VolSplat	0.6951	0.3593	0.6539	0.3219	0.2528	0.2053	0.4740	0.2413	0.4248	0.2277	0.6663	0.3312	0.5278	0.2811
	WorldMirror 2.0	0.4557	0.2009	0.5264	0.2041	0.1003	0.0705	0.4356	0.1974	0.1886	0.1213	0.3520	0.1954	0.3431	0.1649
	Ours	0.3523	0.1259	0.4646	0.1759	0.0634	0.0527	0.3166	0.1387	0.1189	0.0784	0.2226	**0.1436**	0.2564	0.1192
	Ours*	**0.3409**	**0.1215**	**0.4462**	**0.1732**	**0.0579**	**0.0499**	**0.3078**	**0.1351**	0.1134	0.0771	**0.2124**	0.1448	**0.2464**	**0.1169**

Table 6: Per-dataset results with $N = 8$ input views. Mean weights datasets equally.

Input	Method	DL3DV		ETH3D		NRGBD		Mip-NeRF 360		RE10K		DTU		Mean	
		LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓
Pose-free	AnySplat	0.5992	0.2499	0.5586	0.2331	0.1895	0.1128	0.5573	0.2257	0.2486	0.1134	0.4759	0.2332	0.4382	0.1947
	C3G	0.7374	0.3521	0.7020	0.3616	0.5210	0.2536	0.7347	0.3706	0.4404	0.1951	0.6952	0.3710	0.6384	0.3173
	QuerySplat	0.5581	0.2246	0.5953	0.2840	0.2578	0.1394	0.6096	0.2652	0.2148	0.1106	0.5234	0.2737	0.4599	0.2163
	DA3-GS	0.5137	0.2117	0.4897	0.2431	0.3514	0.1794	0.4796	0.2362	0.2592	0.1467	0.4152	0.2337	0.4181	0.2084
	WorldMirror 2.0	0.4796	0.2154	0.4908	0.2317	0.2049	0.1536	0.4701	0.2143	0.2218	0.1334	**0.3690**	0.2074	0.3727	0.1926
	Ours	0.3719	**0.1276**	0.4024	**0.1654**	0.1101	0.0787	**0.3768**	**0.1482**	0.1451	0.0813	0.3782	**0.1910**	0.2974	**0.1320**
	Ours*	**0.3646**	0.1292	**0.3899**	0.1698	**0.1037**	**0.0777**	0.3778	0.1563	**0.1439**	**0.0807**	0.3707	0.2036	**0.2917**	0.1362
Pose-conditioned	tttLRM	0.4910	0.2639	0.5081	0.2385	0.2522	0.1825	0.3617	0.1922	0.2032	0.1338	0.3605	0.2488	0.3628	0.2099
	DA3-GS	0.5156	0.2042	0.5857	0.2505	0.3569	0.1555	0.5018	0.2420	0.2310	0.1264	0.3656	0.2060	0.4261	0.1974
	DepthSplat	0.5897	0.2753	0.5814	0.3046	0.0798	0.0656	0.4595	0.2477	0.1261	0.0815	0.4242	0.2381	0.3768	0.2021
	ReSplat	0.4240	0.2175	0.5472	0.2454	0.0924	0.0887	0.3662	0.1740	0.1214	0.0860	0.4437	0.2732	0.3325	0.1808
	TokenGS	0.6907	0.3120	0.6730	0.2944	0.5410	0.2987	0.6388	0.3032	0.3896	0.2076	0.5043	0.3189	0.5729	0.2891
	VolSplat	0.7064	0.3528	0.6813	0.3165	0.2503	0.2035	0.4964	0.2514	0.3848	0.2121	0.7307	0.3831	0.5416	0.2866
	WorldMirror 2.0	0.4657	0.2135	0.5273	0.2118	0.1061	0.0710	0.4711	0.2186	0.2175	0.1367	0.3265	0.1921	0.3524	0.1740
	Ours	0.3280	0.1191	0.4383	0.1711	0.0656	0.0533	0.3158	0.1374	0.1207	**0.0803**	0.2041	**0.1382**	0.2454	**0.1166**
	Ours*	**0.3191**	**0.1179**	**0.4216**	**0.1669**	**0.0618**	**0.0529**	**0.3068**	**0.1356**	**0.1184**	0.0812	**0.2006**	0.1455	**0.2380**	0.1167

Table 7: Per-dataset results with $N = 16$ input views. Mean weights datasets equally.

Input	Method	DL3DV		ETH3D		NRGBD		Mip-NeRF 360		RE10K		DTU		Mean	
		LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓
Pose-free	AnySplat	0.5958	0.2445	0.5588	0.2315	0.1873	0.1112	0.5705	0.2334	0.2490	0.1120	0.4646	0.2251	0.4377	0.1929
	C3G	0.7571	0.3775	0.7056	0.3921	0.4833	0.2398	0.7973	0.4479	0.4315	0.1813	0.7362	0.3722	0.6518	0.3351
	QuerySplat	0.5769	0.2342	0.6146	0.2986	0.2599	0.1411	0.6510	0.2954	0.2156	0.1116	0.5121	0.2669	0.4717	0.2246
	DA3-GS	0.4953	0.2276	0.5017	0.2581	0.3501	0.1835	0.4483	0.2350	0.2574	0.1441	0.3885	0.2262	0.4069	0.2124
	WorldMirror 2.0	0.4956	0.2269	0.5360	0.2506	0.2457	0.1746	0.4971	0.2408	0.2511	0.1470	0.3821	0.2168	0.4013	0.2095
	Ours	0.3736	**0.1280**	0.4560	**0.1881**	0.1109	0.0799	**0.4000**	**0.1641**	0.1489	0.0838	0.3751	**0.1948**	0.3107	**0.1398**
	Ours*	**0.3673**	0.1318	**0.4444**	0.1919	**0.1055**	**0.0787**	0.4038	0.1799	**0.1480**	**0.0818**	**0.3623**	0.2079	**0.3052**	0.1454
Pose-conditioned	tttLRM	0.3998	0.2123	0.4395	0.2027	0.2554	0.1850	0.3437	0.1837	0.1993	0.1333	0.3392	0.2393	0.3295	0.1927
	DA3-GS	0.5284	0.2274	0.5267	0.2228	0.3531	0.1569	0.5025	0.2550	0.2316	0.1276	0.3220	0.1858	0.4107	0.1959
	DepthSplat	0.5413	0.2646	0.5945	0.3260	0.1163	0.0814	0.5044	0.2915	0.1563	0.0972	0.3640	0.2235	0.3795	0.2140
	ReSplat	0.4043	0.2133	0.4815	0.2034	0.1980	0.1245	0.3627	0.1741	0.1287	0.0877	0.4133	0.2687	0.3314	0.1786
	TokenGS	0.6196	0.2913	0.6618	0.2822	0.5392	0.2984	0.6586	0.3155	0.3838	0.2043	0.4377	0.2969	0.5501	0.2814
	VolSplat	0.7132	0.3495	0.6792	0.3064	0.1996	0.1644	0.5509	0.2826	0.2560	0.1558	0.5301	0.2966	0.4882	0.2592
	WorldMirror 2.0	0.4805	0.2234	0.5066	0.2078	0.1104	0.0735	0.5001	0.2423	0.2461	0.1496	0.3134	0.1914	0.3595	0.1813
	Ours	0.3285	**0.1176**	0.3989	0.1514	0.0669	0.0544	0.3292	0.1453	0.1211	**0.0808**	0.1942	**0.1287**	0.2398	**0.1130**
	Ours*	**0.3227**	0.1196	**0.3850**	**0.1491**	**0.0640**	**0.0543**	**0.3150**	**0.1433**	**0.1210**	0.0823	**0.1911**	0.1362	**0.2331**	0.1141

Table 8: Three-step refinement on ETH3D, averaged over $N = 4, 8, 16$ input views.

Variant	PSNR↑ (dB)	SSIM↑	LPIPS↓	DISTS↓
Ours	16.9643	0.5836	0.4339	0.1661
Ours + Adam	16.9890	0.5864	0.4312	0.1656
Ours*	**17.6074**	**0.6067**	**0.4176**	**0.1631**



Figure 10: Three-step refinement on ETH3D with eight input views. Red boxes highlight the compared floor region.

Table 9: Feed-forward network latency: mean and min–max over five runs.

Input resolution ($W \times H$)	Mean (s)	Min–max (s)
1022×756	5.619	5.617–5.621
756×560	2.533	2.532–2.535
504×378	1.032	1.031–1.034
378×280	0.564	0.563–0.564

opacities, rotations, and colors with centers fixed. Adam uses a learning rate of 0.01. Learned refinement and Adam improve PSNR by 0.64 dB and 0.02 dB, respectively; the learned refiner also performs better on the other three metrics. Figure 10 provides a qualitative comparison.

B.4 Opacity and Geometric Consistency

We analyze feed-forward outputs on held-out Hypersim (Roberts et al., 2021) validation scenes. Geometric labels are assigned independently of opacity using GT depth, with tolerance $\max(0.03\,\mathrm{m}, 0.01d)$ and depth-edge samples excluded. A point is geometrically inconsistent if it violates GT-observed free space in at least two views and agrees with no observed surface depth. Surface-consistent points agree with at least two views and have no free-space violations; others remain uncertain. This identifies 36,125 inconsistent points. Opacity distributions and pruning statistics at threshold 0.01 are shown in Fig. 7.

B.5 Input Resolution and Feed-Forward Network Latency

We measure feed-forward neural-module latency at four resolutions on one H20, using the same ETH3D scene and eight input views. Each resolution uses one warmup followed by five synchronized CUDA-event measurements.

B.6 Additional Qualitative Results

Pose-free comparisons. Figure 11 compares pose-free reconstructions on ETH3D. The selected crops show fewer repetitive stripe artifacts and better-preserved relief structures than several baselines.

Diverse scenes. Figure 12 shows feed-forward reconstructions on DDAD (Guizilini et al., 2020), Waymo (Sun et al., 2020), ETH3D (Schops et al., 2017), Hypersim (Roberts et al., 2021), DTU (Aanæs et al., 2016), Mip-NeRF 360 (Barron et al., 2022), and Mill19 (Mega-NeRF) (Turki et al., 2022), covering streets, buildings, interiors, and objects.

Large-scale reconstruction. We apply the feed-forward reconstructor to a KITTI (Geiger et al., 2013) street sequence of over 1,000 frames using sliding windows to construct a large-scale Gaussian scene (Fig. 13).

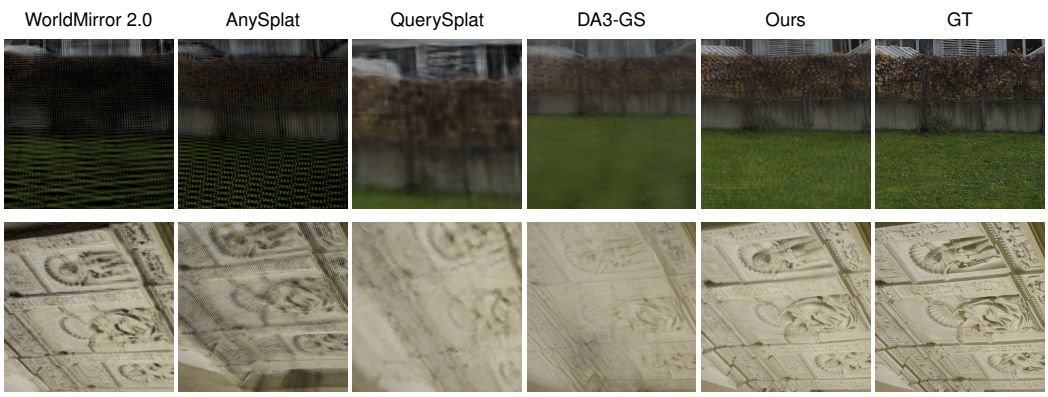

Figure 11: Pose-free rendering on ETH3D: grass textures and relief details.

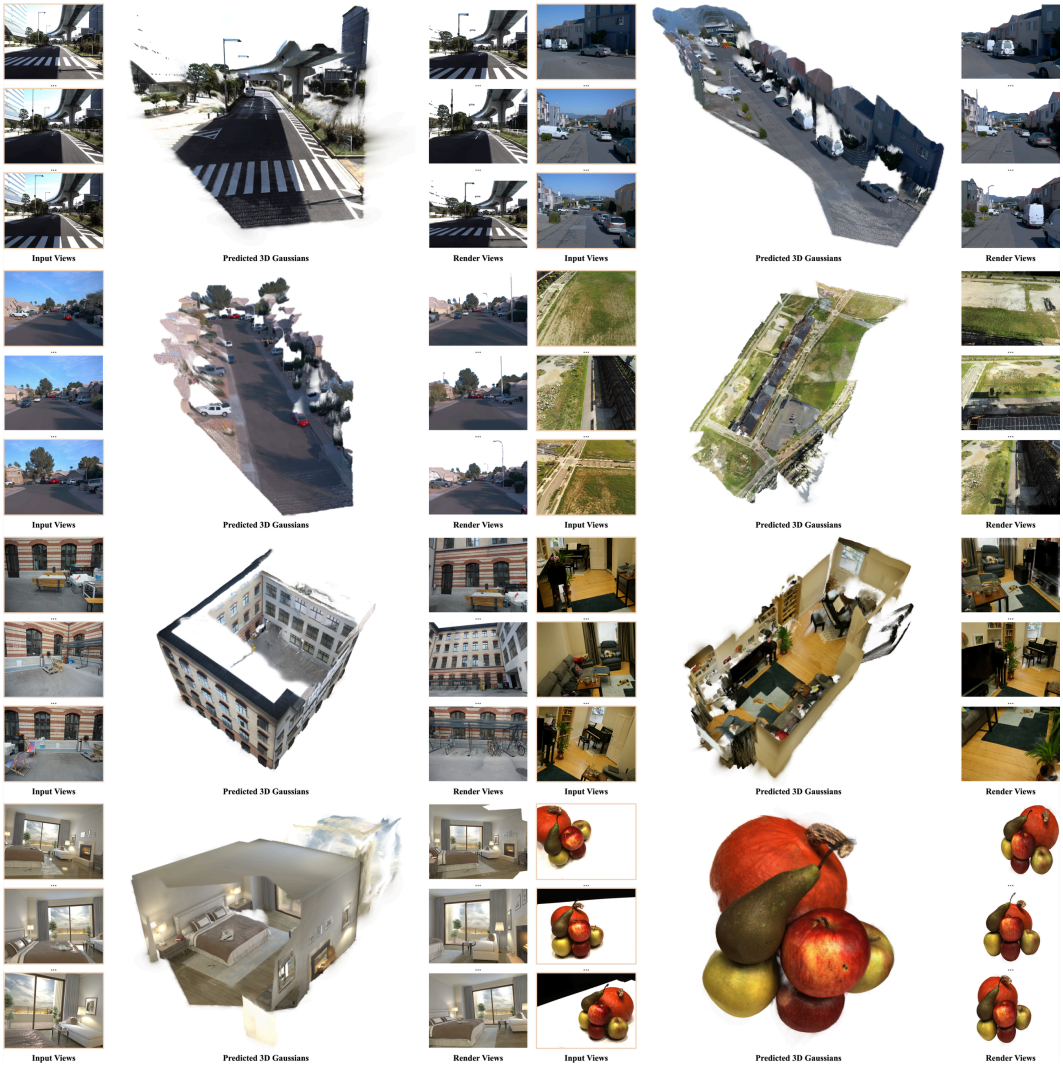

Figure 12: Feed-forward reconstructions and renderings with 12–64 input views.

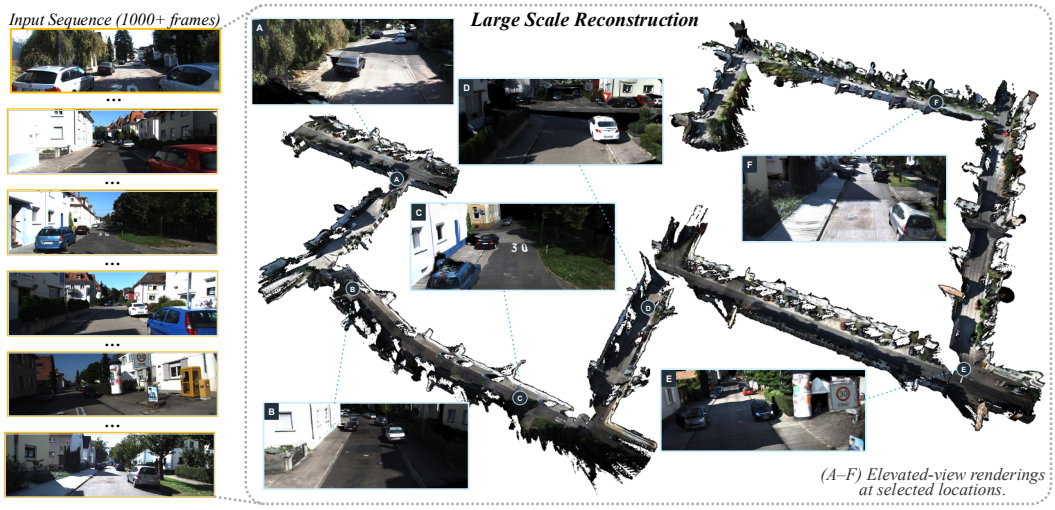

Figure 13: Large-scale reconstruction on KITTI. Insets show elevated-view renderings at locations A–F.

C Limitations and Future Work

Opacity learning and pruning can suppress unreliable Gaussians, but fixed centers cannot correct severe errors in predicted geometry. Reconstruction quality therefore remains limited by the geometric prior. Pixel-aligned prediction also produces redundancy not fully removed by voxel merging, while limited input coverage can leave holes in unobserved regions. Future work will explore stronger geometric priors and more compact Gaussian representations, as well as generative models and camera-trajectory design to address limited coverage.