NOVA3R: NON-PIXEL-ALIGNED VISUAL TRANSFORMER FOR AMODAL 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present NOVA3R, an effective approach for non-pixel-aligned 3D reconstruction from a set of unposed images, in a feed-forward manner. Unlike pixel-aligned methods that tie geometry to per-ray predictions, our formulation learns a global, view-agnostic scene representation that decouples reconstruction from pixel alignment. This addresses two key limitations in pixel-aligned 3D: (1) it recovers both visible and invisible points with a complete scene representation, and (2) it produces physically plausible geometry with fewer duplicated structures in overlapping regions. To achieve this, we introduce a scene-token mechanism that aggregates information across unposed images and a diffusion-based 3D decoder that reconstructs complete, non–pixel-aligned point clouds. Extensive experiments on both scene-level and object-level datasets demonstrate that NOVA3R outperforms state-of-the-art methods in terms of reconstruction accuracy and completeness.

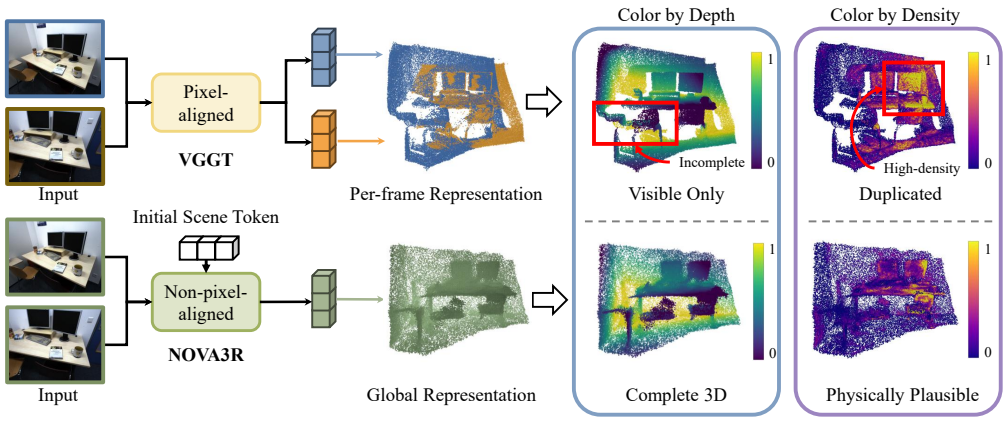

Figure 1: **NOVA3R** enables non–pixel-aligned reconstruction by learning a global scene representation from unposed images. Compared to pixel-aligned methods, NOVA3R recovers both visible and occluded regions and produces more physically plausible geometry with fewer duplicated structures.

1 INTRODUCTION

We consider the problem of *non–pixel-aligned* 3D reconstruction from one or more unposed images, in a feed-forward manner. This is a challenging task, as the model must infer a global, view-agnostic representation of the scene without relying on per-ray supervision. This formulation avoids the limitations of pixel-aligned methods, which reconstruct only visible surfaces and often produce redundant geometry in overlapping regions. It therefore enables more complete and physically plausible 3D reconstruction, capturing both visible and occluded structures in a consistent manner.

Recent work in 3D reconstruction has largely focused on the *pixel-aligned* formulation, where geometry is predicted in the form of depth maps, point maps, or radiance fields tied to the image plane. DUSt3R (Wang et al., 2024a) pioneers this paradigm of dense, pixel-aligned 3D reconstruction from unposed image collections, achieving impressive results in reconstructing the visible regions of a scene. Building on this, follow-up works (Tang et al., 2025b; Wang et al., 2025b; Yang et al., 2025; Zhang et al., 2025b; Wang et al., 2025a) extend DUSt3R from image pairs to multi-view settings, en-

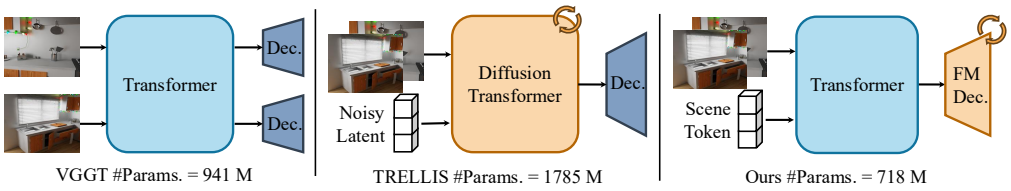

Figure 2: **Comparison of different reconstruction paradigms.** Our non-pixel-aligned approach combines feed-forward efficiency with a global, view-agnostic scene representation, removing the reliance on pixel-level supervision. NOVA3R provides a unified solution for various reconstruction tasks, achieving multi-view consistency and geometrically faithful results.

abling feed-forward 3D geometry reconstruction from larger image sets. However, the pixel-aligned formulation remains tied to per-ray prediction, which restricts reconstruction to visible regions and yields *incomplete* geometry and *overlapping point layers* in areas visible to multiple cameras.

Another line of work explores latent 3D generation, which learns a *global representation* in a compact latent space and decodes it into voxels or meshes (Vahdat et al., 2022; Zhang et al., 2023; 2024b; Ren et al., 2024; Xiang et al., 2025a; Tochilkin et al., 2024; Team, 2024; 2025; Li et al., 2025b). While this global formulation can plausibly complete occluded regions beyond the input views, most approaches remain confined to the *object level*. They assume canonical space and require high-quality mesh supervision, which makes these methods struggle with complex, cluttered scenes. For *scene*-level reconstruction, some methods (Chen et al., 2024; Liu et al., 2024; Gao et al., 2024; Szymanowicz et al., 2025) inpaint unseen regions by synthesizing novel views with pre-trained diffusion models and then post-process to recover geometry. However, such pipelines do not guarantee physically meaningful point clouds.

To overcome these limitations, we introduce the Non-pixel-aligned Visual Transformer (NOVA3R) (see Figure 1). First, we address the challenge of non-pixel-aligned supervision by leveraging a diffusion-based 3D autoencoder. It first compresses complete point clouds into compact latent tokens, and then decodes them back into the original space, supervised with a flow-matching loss that resolves matching ambiguities in unordered point sets. Recent works on 3D autoencoders (Zhang et al., 2023; Xiang et al., 2025a; Team, 2024; Li et al., 2025b) have demonstrated the effectiveness of latent representations, but they are primarily designed for object reconstruction, assuming high-quality meshes for supervision. In contrast, our formulation targets scene-level reconstruction and requires only point clouds derived from meshes or depth maps for supervision, enabling it to capture priors of complete 3D scenes and produce physically coherent geometry without duplicated points.

Second, we tackle the problem of mapping unposed images to a global scene representation. Training such a model directly would require massive amounts of complete scene data and computational resources. To improve generalization, our model is built on a pre-trained image encoder from VGGT (Wang et al., 2025a), augmenting it with learnable scene tokens that aggregate information from arbitrary numbers of views and map them into the latent space of our point decoder. This design enables NOVA3R to support both monocular and multi-view reconstruction, without being restricted to a fixed number of inputs. Despite being trained on relatively small datasets, our model generalizes well to unseen scenes, achieving complete and physically plausible reconstructions.

In summary, our main contributions are as follows: (i) We introduce a unified non-pixel-aligned reconstruction pipeline with minimal assumptions, applicable to both object-level and scene-level completed reconstruction tasks. (ii) We address key limitations of pixel-aligned methods, which often produce incomplete point clouds, duplicated geometry, and 3D inconsistencies in overlapping regions. By contrast, our non-pixel-aligned formulation naturally yields complete and evenly distributed geometry. (iii) We integrate a feed-forward transformer architecture with a lightweight flow-matching decoder, effectively bridging the gap between pixel-aligned reconstruction and latent 3D generation, combining feed-forward efficiency with strong 3D modeling capability (see Figure 2).

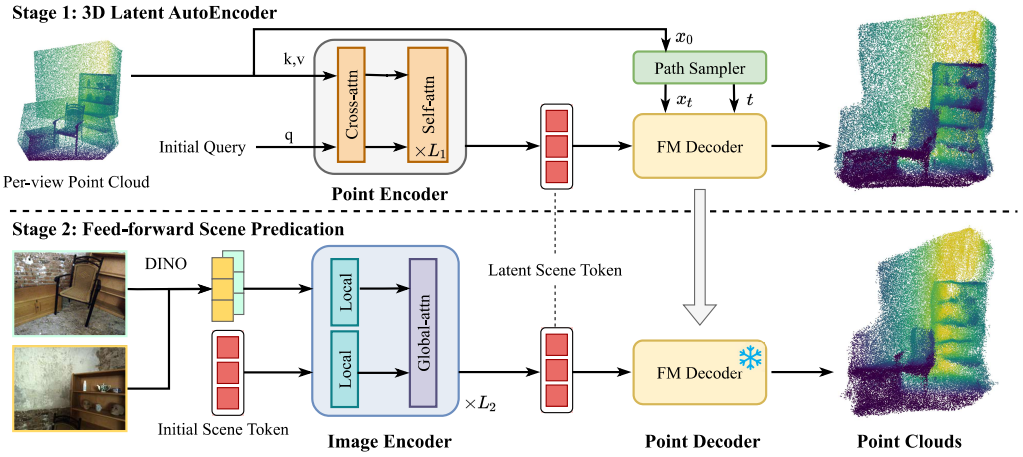

Figure 3: **Overview of NOVA3R. Stage 1:** a 3D point autoencoder encodes complete point clouds into latent scene tokens and decodes them with a flow-matching (FM) decoder. **Stage 2:** an image encoder with learnable scene tokens integrates multi-view information into a unified scene latent space, supervised by the FM loss with the Stage-1 decoder frozen. During **inference**, only the second stage pipeline is used to produce a complete, non–pixel-aligned point cloud.

2 Related Work

Feed-Forward 3D Reconstruction. Unlike *per-scene* optimisation methods (Mildenhall et al., 2020; Kerbl et al., 2023) that iteratively refine a 3D representation for each individual scene, *feed-forward* 3D reconstruction approaches aim to generalize across scenes by predicting 3D geometry directly from a set of input images in a single pass of a neural network. Early approaches typically focus on predicting geometric representations, such as depth maps (Eigen & Fergus, 2015), meshes (Wang et al., 2018), point clouds (Fan et al., 2017), or voxel grids (Choy et al., 2016), and are trained on relatively small-scale datasets (Nathan Silberman & Fergus, 2012; Chang et al., 2015). As a result, these models struggled to capture fine-grained visual appearance and exhibited limited generalization to unseen scenes.

More recently, DUSt3R (Wang et al., 2024a) and MASt3R (Leroy et al., 2024) directly regress dense, pixel-aligned point maps from unposed image collections. These approaches mark a significant step toward generalizable, pose-free 3D reconstruction. Building on this diagram, many recent works (Tang et al., 2025b; Wang et al., 2025b; Yang et al., 2025; Zhang et al., 2025b; Wang et al., 2025a) extend it from image pairs to multi-view settings, enabling feed-forward 3D geometry reconstruction from sets of uncalibrated images. However, these pixel-aligned methods produce incomplete geometry and duplicated points in overlapping regions. In contrast, our approach outputs a unified and *complete* 3D reconstruction that integrates both *visible* and *occluded* regions.

Complete 3D Reconstruction. To achieve a complete 3D reconstruction, existing approaches typically follow two main paradigms. One line of work (Vahdat et al., 2022; Zhang et al., 2023; Zhao et al., 2023; Zhang et al., 2024b; Ren et al., 2024; Xiang et al., 2025a; Tochilkin et al., 2024; Team, 2024; 2025; Li et al., 2025b) leverages compact latent spaces (Rombach et al., 2022) or large-scale networks (Hong et al., 2024; Zhang et al., 2024a; Tang et al., 2025a) for generating complete 3D assets. While effective, these approaches primarily target individual *object* reconstruction and fall short in modeling complex, cluttered scenes. The other paradigm fine-tunes large-scale pre-trained diffusion models (Rombach et al., 2022; Blattmann et al., 2023). For *objects*, a notable example is Zero-1-to-3 (Liu et al., 2023b), which conditions on camera pose for high-quality 360-degree novel view rendering by training on a huge dataset, Objaverse (Deitke et al., 2023). This is followed by a large group of successors (Long et al., 2024; Shi et al., 2024; Han et al., 2024; Liu et al., 2023a; Li et al., 2024; Zheng & Vedaldi, 2024; Ye et al., 2024; Voleti et al., 2024). For *scenes*, several recent approaches aim to achieve complete 3D geometry by leveraging controlled camera trajectories (Wang et al., 2024b; Sargent et al., 2024; Wu et al., 2024; Gao et al., 2024; Wallingford et al., 2024; Zhou et al., 2025) or introducing auxiliary conditioning signals (Liu et al., 2024; Yu et al., 2024; Chen et al., 2024; Yu et al., 2025). However, these methods do not explicitly re-

construct the complete underlying 3D geometry. More recently, WVD (Zhang et al., 2025a) and Bolt3D (Szymanowicz et al., 2025) propose a hybrid RGB+point map representation to combine geometry and appearance for 3D reconstruction; however, they still require known camera poses for novel RGB+point map rendering. We address *pose-free* 3D reconstruction from unconstrained images, and provide a complete 3D representation. More closely related to our work, Amodal3R (Wu et al., 2025) introduces amodal 3D reconstruction to reconstruct complete 3D assets from partially visible pixels, but it still works only on objects.

3 METHOD

Given a set of unposed images $\mathcal{I} = \{\boldsymbol{I}^i\}_{i=1}^K$, ($\boldsymbol{I}^i \in \mathbb{R}^{H\times W\times 3}$) of a scene, our goal is to learn a neural network Φ that directly produces a complete 3D point cloud, both in terms of *visible* and *occluded* regions. We first discuss the problem formulation in section 3.1, followed by our 3D latent autoencoder in section 3.2, and we finally describe our global scene representation in section 3.3.

3.1 PROBLEM FORMULATION

Problem Definition. The input to our model is a set of K *unposed images* $\mathcal{I} = \{\boldsymbol{I}^i\}_{i=1}^K$ of a scene, and the output is a *complete* 3D point cloud $P \in \mathbb{R}^{N\times 3}$, using a feed-forward neural network $\Phi : \mathcal{I} \rightarrow P$. This task is conceptually similar to the conventional feed-forward 3D reconstruction setting (Wang et al., 2024a; 2025a; Jiang et al., 2025), except that here N represents the number of points in the *complete* scene point cloud (as shown in fig. 4), rather than $K \times H \times W$ points back-projected from all pixels in each input image.

The *key observation* is that a scene in the real world is composed of a fixed set of physical points, regardless of how many images are captured from different viewpoints. If a physical 3D point is observed in multiple 2D images, the correct representation of the scene should contain a single point, rather than duplicated points back-projected from each observation. Conversely, even if a physical 3D point is never observed in any image, it still exists in the real world and should be inferred by the model. Therefore, the model should be able to predict the occluded regions of the scene and avoid generating redundant points in the overlapping visible regions.

Data Preprocessing. The key to training such a model is the definition of the *complete* 3D point clouds of a scene. It must contain points in both *visible* and *occluded* regions, and avoid duplicated points in the *overlapping visible* regions. The visibility of a 3D point is defined with respect to the input images $\mathcal{I}$. However, the notion of invisible points is ambiguous: there are infinitely many points that are not visible in input images, or even outside the field of view of all input images. To simplify the problem, as shown in Figure 4, we define invisible points within the input-view frustum and discard points outside the frustum.

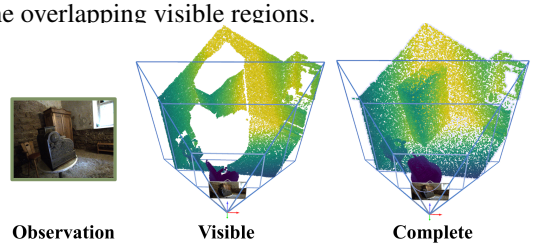

Figure 4: **Visible point clouds *vs.* complete point clouds.** Our NOVA3R aims to recover the complete geometry within the input view's frustum.

Creating such complete point clouds for supervision is non-trivial. The ideal solution is to use the ground-truth 3D mesh of the scene, which can be easily converted to a complete point cloud by uniformly sampling points on the mesh surface. However, the ground-truth mesh is not always available in scene-level datasets. When ground-truth 3D meshes are not available, we instead approximate the complete point clouds using depth maps aggregated from dense views. Specifically, we first back-project the depth maps from all dense views into point clouds, then apply voxel-grid filtering to remove duplicate points in overlapping visible regions. Finally, we discard points outside the frustum of the selected input views (single, two, or a set of views). During training, we apply the farthest point sampling method with random initialization to obtain a subset from the complete point cloud to train our point decoder.

Importantly, as in DUSt3R (Wang et al., 2024a), our completed point clouds are also *view-agnostic*: the 3D points are defined in the coordinate system of the first input view $\boldsymbol{I}_1$, but are *not* pixel-aligned to any input images. This design allows the model to learn to reconstruct the complete 3D scene in

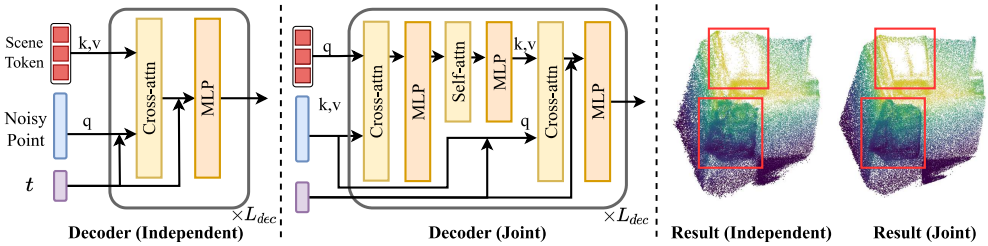

Figure 5: **Different Decoder Architectures.** The independent decoder uses cross-attention only, while the joint decoder implements an efficient self-attention, which yields more precise structures.

the first view's coordinate system while ignoring the ambiguity of pose estimation. Consequently, our model can be trained on a wide range of datasets without requiring ground-truth meshes, unlike existing object-level methods (Zhang et al., 2023; Li et al., 2025b; Team, 2024).

3.2 3D Latent Encoder-Decoder with Flow Matching

Following recent works in 3D latent vector representation (Zhang et al., 2023), we design a 3D latent transformer (Vaswani et al., 2017). However, ours does *not* require a perfect mesh as input or supervision. As shown in Figure 3 (Stage 1), we implement the model as a diffusion model.

Diffusion-based 3D AutoEncoder. The encoder Φ_{enc} takes the point cloud $P \in \mathbb{R}^{N\times 3}$ as input, and outputs a set of M latent tokens $Z \in \mathbb{R}^{M\times C}$. In practice, to reduce the computational cost, the initial query points $q \in \mathbb{R}^{M\times 3}$ are sampled from the complete point cloud $P \in \mathcal{R}^{N\times 3}$ using farthest point sampling, where $M \ll N$. We further design a hybrid query representation by concatenating the point query with learnable tokens of the same length M along the channel dimension, followed by a linear projection layer that reduces the channel dimension from $2C$ to C.

Once the latent tokens Z are obtained, existing 3D VAE methods (Zhang et al., 2023; Team, 2025; Li et al., 2025b) typically use a deterministic decoder to predict an occupancy field $o = \Phi_{\text{dec}}(Z, x)$ or SDF values $s = \Phi_{\text{dec}}(Z, x)$ for each 3D grid query $x \in \mathbb{R}^{N\times 3}$. However, this is not suitable for our task, since obtaining ground-truth occupancy or SDF values for real scene-level datasets is costly or even infeasible. Importantly, unlike objects that can be enclosed within a canonical space, scenes typically lack well-defined boundaries and expand as the number of observations increases, making it difficult to predefine a canonical space. Instead, we directly predict the 3D coordinates of each query point. However, because point clouds are *not* ordered or aligned, we cannot directly map the 3D point query to the ground-truth point clouds P using an L_2 loss. We then adopt a diffusion-based decoder $\Phi_{\text{dec}}(x_t, Z, t)$ to decode the scene tokens Z back to the original point cloud space. The transformer-based decoder takes as input a set of N noised query point clouds $x_t \in \mathbb{R}^{N\times 3}$, at the flow matching time t, and the latent tokens Z as conditioning. The whole architecture is trained end-to-end with a flow matching loss (Lipman et al., 2023):

$$\mathcal{L}^{\text{AE}}_{\text{flow}} = \mathbb{E}_{t,x_0\sim P,\epsilon\sim\mathcal{U}(-1,1)}\left[\|\Phi_{\text{dec}}(x_t, Z, t) - (\epsilon - x_0)\|_2^2\right], \tag{1}$$

where $x_t = (1-t)x_0 + t\epsilon$. Note that, we do *not* use KL loss or any other regularization on the latent tokens as in existing 3D latent VAE methods (Team, 2025; Li et al., 2025b).

Architecture. As noted above, our 3D autoencoder is implemented with a transformer architecture. Specifically, the encoder is built upon TripoSG (Li et al., 2025b), which consists of one cross-attention layer and eight self-attention layers. The decoder has three transformer blocks (details are shown in Figure 5). Notably, the query will be switched between the 3D latent tokens Z and the noisy point clouds x_t in each cross-attention layer. This design reduces the size of the self-attention maps while preserving information flow between latent tokens and query points. Concurrent work (Chang et al., 2024) also proposes a diffusion-based 3D latent autoencoder, but they consider a 3D shape as a probability density function, and process each point independently.

3.3 Scene Representation with Learnable Tokens

We now describe how to learn a global scene representation from a set of unposed images. As shown in Figure 3 (Stage 2), we implement it using a large transformer that takes the input images $\mathcal{I}$ and a set of M learnable tokens $t_S \in \mathbb{R}^{M\times C}$ as input, and outputs the scene representation $\hat{Z} \in \mathbb{R}^{M\times C}$.

Table 1: **Quantitative results for scene completion on SCRREAM** (Jung et al., 2024). The *one-side* Chamfer Distance (GT → Prediction) results are shown in ( ). K is the number of input views. $*$ denotes methods that are not trained on scene-level data. Our method shows better completion results compared to other competitive baselines. Note that, since NOVA3R is a *non-pixel-aligned* 3D reconstruction model, it does not explicitly distinguish the visible and occluded points.

Type	Method	Visible (K=1)			Complete (K=1)			Complete (K=2)		
		CD↓	FS@0.1↑	FS@0.05↑	CD↓	FS@0.1↑	FS@0.05↑	CD↓	FS@0.1↑	FS@0.05↑
Object	TripoSG*	(0.268)	(0.418)	(0.301)	0.242	0.467	0.333	-	-	-
	TRELLIS*	(0.301)	(0.420)	(0.313)	0.256	0.429	0.312	0.286	0.402	0.288
Single-view	Metric3D-v2	0.063	0.803	0.534	0.086	0.725	0.473	-	-	-
	DepthPro	0.055	0.852	0.603	0.079	0.764	0.535	-	-	-
	MoGe	**0.035**	**0.945**	**0.786**	0.063	0.836	0.668	-	-	-
	LaRI	0.057	0.847	0.589	0.059	0.825	0.590	-	-	-
Multi-view	DUST3R	0.059	0.851	0.653	0.086	0.757	0.565	0.061	0.833	0.641
	CUT3R	0.069	0.835	0.679	0.091	0.753	0.543	0.092	0.739	0.532
	VGGT	0.041	0.923	0.754	0.070	0.810	0.657	0.065	0.821	0.606
	Ours	(0.043)	(0.904)	(0.730)	**0.048**	**0.882**	**0.687**	**0.053**	**0.862**	**0.657**

Learnable Scene Tokens. As mentioned in Section 3.1, our model aims to predict a *fixed* number of *non-pixel-aligned* point cloud undelying the first view's coordinate system. Accordingly, in addition to L patchified image tokens $t_I \in \mathbb{R}^{L\times C}$, we introduce a set of M learnable global scene tokens $t_S \in \mathbb{R}^{M\times C}$, which are randomly initialized and optimized during training. The combined token set $t_I \cup t_s$ from all input images, *i.e.*, $t_I = \cup_{i=1}^{K}\{t_I^i\}$, and the learnable scene tokens t_S, is fed into a large transformer, with multiple frame- and global-level self-attention layers. To simplify the architecture, the learnable scene tokens t_S are treated as a global frame underlying the first view's coordinate system. This means that the scene tokens undergo the same operations as the image tokens in each Transformer block, except that they use the first view's camera token.

Architecture. Our image encoder is built upon VGGT (Wang et al., 2025a), a representative feed-forward 3D reconstruction model. However, we do not use its dense predication heads to predict the *pixel-aligned* depth and point maps. Instead, we use the output scene tokens $\hat{Z} \in \mathbb{R}^{M\times C}$ as the conditioning of our point decoder Φ_{dec}, to predict the *non-pixel-aligned* complete 3D point clouds $\hat{P} \in \mathbb{R}^{N\times 3}$. The entire architecture is trained end-to-end with the flow matching loss:

$$\mathcal{L}_{\text{flow}}^{\text{Tran}} = \mathbb{E}_{t,x_0\sim P,\epsilon\sim\mathcal{U}(-1,1)}\left[\left\|\Phi_{\text{dec}}(x_t,\hat{Z},t)-(\epsilon-x_0)\right\|_2^2\right], \tag{2}$$

where Φ_{dec} is frozen in Stage 2, and only the transformer $\Phi_{\text{tran}} : t_I \cup t_S \to \hat{Z}$ and the learnable scene tokens t_S are optimized.

4 Experiments

4.1 Experimental Settings

Metrics. Following Li et al. (2025a), we report Chamfer Distance (CD) and F-score (FS) at different thresholds (*e.g.,* 0.1, 0.05) for completion tasks. For multi-view reconstruction tasks, we report accuracy (Acc), completion (Comp), and normal consistency (NC) following Wang et al. (2025b). Best results are highlighted as **first**, second, and third.

Implementation Details. By default, we set the number of scene tokens as $M = 768$ and the number of points as $N = 10,000$ for training. The image encoder architecture is exactly the same as VGGT (Wang et al., 2025a), while the 3D latent autoencoder contains 8 layers in the encoder and 3 layers in the decoder. The training contains two stages. In Stage 1, we train the autoencoder for 50 epochs. In Stage 2, we initialize the image encoder with VGGT pretrained weights and the flow-matching decoder with Stage-1 weights, then train for another 50 epochs. Note that, we only fine-tune the image encoder and the scene-token transformer in Stage 2. We train both stages by optimizing the flow-matching loss with the AdamW optimizer and a learning rate of 3e-4. The training runs on 4 NVIDIA A40 GPUs with a total batch size of 32. We use standard flow-matching with cosine noise scheduling, timestep sampling in [0,1], a fixed 0.04 step size at inference, and identical loss settings for both object-level and scene-level datasets.

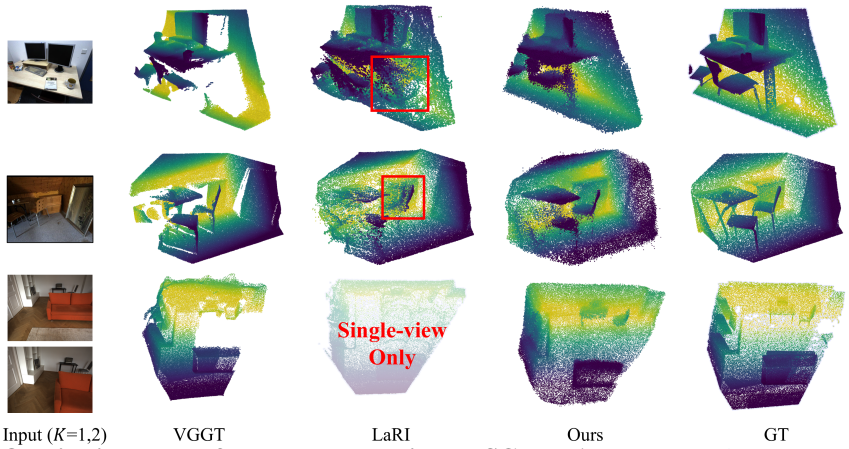

Figure 6: **Qualitative results for scene completion on SCRREAM** (Jung et al., 2024). Our method produces more complete point clouds with clearer and less distorted geometry than other baselines.

Table 2: **Quantitative results for hole area ratio and point cloud density variance on SCRREAM (Jung et al., 2024).** Our method significantly outperforms pixel-aligned baselines, achieving both lower hole ratios and lower density variance.

Method	**Complete** (K=1)		**Complete** (K=2)		**Complete** (K=4)	
	Hole Ratio↓	Density Var. ↓	Hole Ratio ↓	Density Var. ↓	Hole Ratio ↓	Density Var. ↓
DUST3R	0.317	7.758	0.237	6.553	0.257	4.801
CUT3R	0.363	8.402	0.237	6.554	0.326	4.658
VGGT	0.307	7.105	0.238	6.546	0.261	5.217
Ours	**0.088**	**5.127**	**0.121**	**2.188**	**0.134**	**1.881**

4.2 Scene-level Reconstruction

Datasets. The scene-level model was trained on 3D-FRONT (Fu et al., 2021) and ScanNet++V2 (Yeshwanth et al., 2023), using the training splits from LaRI (Li et al., 2025a) and DUSt3R (Wang et al., 2024a), which contain 100k and 230k unique images, respectively. For visible part training, we further incorporate ARKitScenes (Baruch et al., 2021). Ideally, our model is able to handle an arbitrary number of input views, similar to VGGT (Wang et al., 2025a). However, limited by the available computational resources, we mainly verify our contributions on two-view pairs and train with 1–2 input views.

To evaluate the cross-domain generalisation ability of our model, we directly evaluate performance on the unseen SCRREAM dataset (Jung et al., 2024), which provides complete ground-truth scans. We follow LaRI's setting for single-view evaluation, with 460 images for testing. For the two-view setting, we sample 329 pairs from the same scene with a frame-ID distance of 40–80, where the maximum pose gap is 30% (measured by point cloud covisibility) and the hole area ratios (measured by completeness with threshold 0.1) range from 5.3% to 48.6%. We additionally evaluate visible-surface multi-view reconstruction on the 7-Scenes (Shotton et al., 2013) and NRGBD datasets (Azinović et al., 2022), sampling input images at intervals of 100 frames.

Baselines. We compare NOVA3R with several representative scene-level 3D reconstruction methods, including **i)** single-view Metric3D-v2 (Hu et al., 2024), DepthPro (Bochkovskii et al., 2024), and MoGe (Wang et al., 2025c); **ii)** multi-view DUSt3R (Wang et al., 2024a), CUT3R (Wang et al., 2025b), and VGGT (Wang et al., 2025a). However, these methods only focus on *pixel-aligned visible* 3D reconstruction. Hence, we further compare with the concurrent complete 3D reconstruction work LaRI (Li et al., 2025a). Since it does not support multi-view inputs, for completeness, we also report the results from object-level methods TripoSG (Li et al., 2025b) and TRELLIS (Xiang et al., 2025b) by disabling the input mask, while they are not trained on scene-level data.

Scene Completion. Following LaRI, we evaluate our amodal 3D reconstruction results on both *visble* and *complete (visible + occluded)* regions. For visible setting, we follow the same evaluation protocol as DUST3R (Wang et al., 2024a) and VGGT (Wang et al., 2025a), where the ground

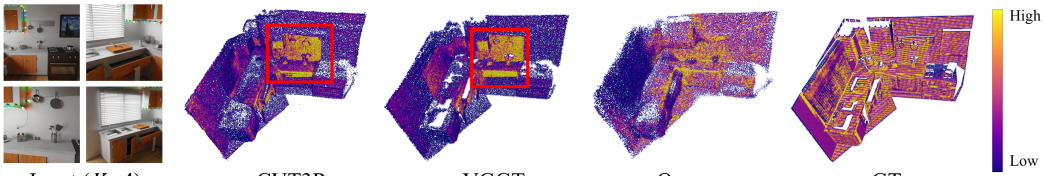

Figure 7: **Qualitative results for density evaluation on NRGBD** ($K = 4$) **(Azinović et al., 2022).** Yellow regions denote higher density, and purple regions denote lower density. Despite being trained with only two views, NOVA3R generalizes well to multiple views ($K = 4$).)

Table 3: **Quantitative results on visible reconstruction on 7-Scenes** (K**=2) (Shotton et al., 2013).** Our NOVA3R model can be trained on RGB-D data and achieves competitive results compared to multi-view reconstruction methods. Note that, we use much less token to represent a 3D scene.

Method	**# Tokens**	Acc ↓		Comp ↓		NC ↑	
		Mean	Med.	Mean	Med.	Mean	Med.
DUSt3R	2048	0.054	0.023	0.075	0.034	0.772	0.901
Spann3R	784	0.044	0.022	0.046	0.025	0.792	0.922
CUT3R	768	0.043	0.023	0.054	0.028	0.760	0.884
VGGT	2738	0.042	**0.020**	0.045	0.025	**0.813**	**0.923**
Ours	768	**0.041**	0.021	**0.033**	**0.019**	0.794	0.917

truth contains only the visible points from the input views. For the complete setting, we use the full point cloud as ground truth, including occluded and unseen regions. However, unlike pixel ray-conditional LaRI, NOVA3R does not explicitly identify the visible region. We therefore adopt *one-sided* Chamfer Distance (GT → Prediction) for the visible part: each GT-visible point must be explained by a nearby prediction. This measures coverage of the visible ground truth, yet without penalizing missing, occluded regions. Table 1 shows three settings: 1-view *visible*, 1-view *complete*, and 2-view *complete*. Despite using only two datasets to train, our method consistently outperforms multi-view baselines on complete reconstruction in both $K = 1$ and $K = 2$ settings, demonstrating the effectiveness of our non–pixel-aligned approach. Our method also achieves competitive results on visible-surface reconstruction. Qualitative results in Figure 6 show that our method produces surfaces without holes (unlike pixel-aligned methods such as VGGT) and yields clearer, less distorted geometry than LaRI. Such benefit is attributed to our non–pixel-aligned design, which prevents ray-direction bias in reconstruction. We further quantify the completion capability using the hole area ratio, which is computed by checking whether each ground-truth point has a predicted point within a distance threshold of 0.1. As shown in Table 2, our method consistently achieves significantly lower hole ratios, demonstrating its strong capability for complete reconstruction. In terms of density variance, our approach outperforms all pixel-aligned baselines, even in unseen four-view settings, indicating better physical plausibility with more evenly distributed point clouds. Moreover, when comparing across different K, the density variance consistently decreases from one to four input views, further confirming that incorporating more views leads to improved spatial uniformity.

Physically-reasonable Reconstruction. Except for 3D completion, our *non–pixel-aligned* formulation also features physically plausible reconstruction by fusing evidence in 3D rather than along camera-pixel rays, reducing duplicated points in overlapping regions and improving cross-view consistency. To illustrate this, we evaluate visible reconstruction with $K = 4$ views on NRGBD (Azinović et al., 2022). As shown in Figure 7, pixel-aligned methods like CUT3R (Wang et al., 2025b) and VGGT (Wang et al., 2025a) accumulate 3D points in co-visible regions, producing uneven densities and multi-layer artifacts. This is physically incorrect, as each point corresponds to a single location in the real world, regardless of the number of views. In contrast, our NOVA3R generates cleaner, single-surface geometry with uniform point distribution, achieving competitive results despite using fewer datasets and views (see Table 3). We further quantify physical plausibility by computing the density variance in Table 2, which indicates that our method achieves a more uniformly distributed reconstruction compared to pixel-aligned baselines.

4.3 Object-Level Reconstruction

Datasets. We demonstrate the versatility of our method as a unified non–pixel-aligned approach for both scenes and objects. Following Li et al. (2025a), we train an object-completion model on

Table 4: **Quantitative results for object completion on GSO (Downs et al., 2022).** NOVA3R provides a unified solution for both scene and object completion with unposed images input.

Type	Method	View-aligned (K=1)			View-aligned (K=2)		
		CD ↓	FS@0.1 ↑	FS@0.05 ↑	CD ↓	FS@0.1 ↑	FS@0.05 ↑
Single-view	SF3D	0.037	0.913	0.738	-	-	-
	SPAR3D	0.038	0.912	0.745	-	-	-
	LaRI	0.025	0.966	0.894	-	-	-
	TripoSG	0.025	0.961	0.899	-	-	-
Multi-view	TRELLIS	0.025	0.962	0.896	0.028	0.946	0.874
	Ours	**0.020**	**0.985**	**0.925**	**0.023**	**0.978**	**0.903**

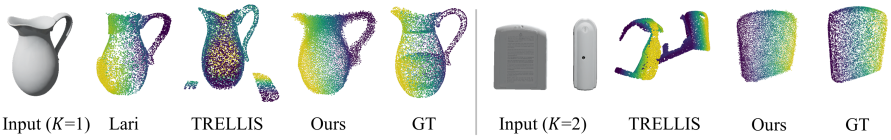

Figure 8: **Qualitative results for object completion on GSO (Downs et al., 2022).** Our method provides more precise geometry and better 3D consistency with multi-view inputs.

Objaverse (Deitke et al., 2023) with 190k annotated images. For evaluation, we report results on unseen Google Scanned Objects (Downs et al., 2022). For single-view reconstruction, we use the same 1030-object split as LaRI (Li et al., 2025a). For two-view reconstruction, we fix the 0th image and uniformly sample three additional views, yielding three pairs per object (3090 pairs in total).

Baselines. We compare with several representative object-level 3D reconstruction methods, including SF3D (Boss et al., 2025), SPAR3D (Huang et al., 2025), TripoSG (Li et al., 2025b), and TRELLIS (Xiang et al., 2025b). We also include LaRI (Li et al., 2025a) as a strong baseline, which is trained on the same dataset and supports amodal 3D reconstruction.

Object Completion. Table 4 reports results for single view ($K = 1$) and two views ($K = 2$). Our NOVA3R outperforms LaRI on all three metrics. Importantly, our pipeline supports multi-view completion that maps different unposed images into the same view-aligned space. On the multi-view benchmark, our method also outperforms TRELLIS, highlighting the benefits of non–pixel-aligned reconstruction for consistent global geometry. Qualitative comparisons in Figure 8 show that our completions preserve fine structures, and achieve better 3D consistency in the multi-view setting.

4.4 Ablation Studies

We perform comprehensive ablation studies on the SCRREAM complete ($K = 1$) setting to validate the key design choices of our method, with particular emphasis on assessing the contribution of Scene Tokens to global structure modeling. The results are summarized in Table 5, and we discuss each component in detail below.

Initial Query (Stage 1). Prior work (Zhang et al., 2023) shows that the initialization of point queries affects autoencoder performance. We compare three options: (i) *downsampled input points*, (ii) *learnable query tokens*, and (iii) a *hybrid* that concatenates (i) and (ii). Downsampled points preserve the input geometry distribution, whereas learnable tokens add flexibility under source–target shifts. As shown in Table 5, the hybrid combines these benefits and yields the best results.

Number of latent scene tokens (Stage 1). As described in Section 3, we represent each scene with a fixed-length set of latent tokens. The number of tokens M directly affects the representation capacity and ability to capture fine details, especially in large scenes. We evaluate different numbers of scene tokens from $\{256, 512, 768\}$ and observe consistent improvements as the count increases (see Table 5). To balance accuracy and efficiency, we use $M = 768$ tokens by default. Ideally, M could be further increased for better performance. We leave this for other works to explore.

Different architecture of flow-matching decoder (Stage 1). The latest work (Chang et al., 2024) also presents a flow-matching decoder for ponint cloud encoder, but it assumes that all points are independent (shown in Figure 5). This design is efficient, but ignores spatial correlations between points. In our work, we instead adopt a lightweight *self-attention* + *cross-attention* decoder that

Table 5: **Ablations.** All models are evaluated on the SCRREAM complete ($K = 1$) setting. We report CD↓, FS@0.1↑, FS@0.05↑ and FS@0.02↑ across different ablation settings.

	Init Query tokens (Stage 1)			# Scene tokens (Stage 1)			FM Decoder (Stage 1)		Img Resolution (Stage 2)	
Settings	Point	Learnable	**Hybrid**	256	512	**768**	Indep.	**Joint**	224	**518**
CD↓	0.011	0.013	**0.011**	0.014	0.013	**0.011**	0.012	**0.011**	0.054	**0.048**
FS@0.1↑	0.999	0.998	**0.999**	0.996	0.998	**0.999**	0.998	**0.999**	0.861	**0.882**
FS@0.05↑	0.991	0.981	**0.993**	0.975	0.986	**0.993**	0.990	**0.993**	0.648	**0.687**
FS@0.02↑	0.894	0.841	**0.904**	0.811	0.839	**0.904**	0.889	**0.904**	0.327	**0.350**

Table 6: **Ablations on different training loss functions.** All models are evaluated on the SCRREAM complete ($K = 1$) setting. We report CD↓, FS@0.1↑, FS@0.05↑ and FS@0.02↑ and inference time↓ for the decoder.

Training Loss	**SCRREAM (Stage 1)**				
	CD ↓	FS@0.1 ↑	FS@0.05 ↑	FS@0.02 ↑	Inference Time (s) ↓
Chamfer distance	0.024	0.981	0.907	0.575	**0.557**
Flow-matching	**0.011**	**0.999**	**0.993**	**0.904**	2.985

jointly reasons over points and scene tokens, allowing information exchange across the point set. To investigate the effect of this design, we compare it with an independent variant without self-attention. Empirically, the joint decoder yields lower CD, higher F-scores, and sharper fine details (Table 5), with small quantitative but significant qualitative gains (Figure 5).

Input image resolution (Stage 2). In Stage 2 (image-to-point), we adopt a transformer to integrate information between image and scene tokens. The input resolution determines the number of image tokens used in the aggregation process. With patch size 14, a resolution of 224×224 yields $16 \times 16 = 256$ tokens, while a resolution of 518×518 yields $37 \times 37 = 1369$ tokens. As shown in Table 5, training with resolution 518 inputs consistently improves CD and F-scores.

Flow-matching loss vs. Chamfer distance loss. To verify the necessity of flow-matching for unordered point cloud encoding, we conducted an ablation using the same architecture but replaced the flow-matching loss with Chamfer Distance. Both models were trained on SCRREAM (Stage 1) under the same protocol. As shown in Table 6, flow-matching achieves significantly better reconstruction quality and generalization. Chamfer Distance struggles in scene-level settings because its nearest-neighbor formulation is computationally expensive, sensitive to density imbalance, and unable to capture global structure across varying scales and input views, while flow-matching produces stable, complete, and globally consistent reconstructions.

5 Conclusion

We present NOVA3R, a non-pixel-aligned framework for amodal 3D scene reconstruction from unposed images. Unlike prior pixel-aligned methods, our NOVA3R achieves state-of-the-art results in amodal, including both visible and invisible points, 3D reconstruction on both scene and object levels. Notably, it also pioneers a paradigm for physically plausible scene reconstruction that reconstructs a uniform point cloud for the entire scene, without holes or duplicated points. This simple yet effective design makes it a promising solution for real-world applications.

Limitations and Discussion. Limited by the computational resources, we train our model with a relatively small number of scene tokens (768) and point clouds (10,000) and a moderate number of input views (up to 2). Hence, the reconstruction quality may degrade for large-scale scenes with complex structures. Future work can explore scaling up the model and training data to enhance performance and generalization. In addition, our model currently focuses on reconstructing static scenes and does not handle dynamic objects or temporal consistency across frames.

References

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, pp. 6290–6301, 2022. 7, 8, 16, 18

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=tjZjv_qh_CE. 7

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

Aleksei Bochkovskii, AmaÃĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 7

Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16240–16250, 2025. 9

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 17

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

Jen-Hao Rick Chang, Yuyang Wang, Miguel Angel Bautista Martin, Jiatao Gu, Xiaoming Zhao, Josh Susskind, and Oncel Tuzel. 3d shape tokenization via latent flow matching. *arXiv preprint arXiv:2412.15618*, 2024. 5, 9

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. In *Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3

Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision (ECCV)*, pp. 628–644. Springer, 2016. 3

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 13142–13153, 2023. 3, 9

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022. 9

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pp. 2650–2658, 2015. 3

Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 605–613, 2017. 3

Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J. Black, Trevor Darrell, and Angjoo Kanazawa. St4rtrack: Simultaneous 4d reconstruction and tracking in the world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 19

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10933–10942, 2021. 7

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3

Junlin Han, Filippos Kokkinos, and Philip Torr. Vfusion3d: Learning scalable 3d generative models from video diffusion models. In *European Conference on Computer Vision (ECCV)*, pp. 333–350. Springer, 2024. 3

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 (ICLR)*, 2024. 3

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 7

Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M Rehg, and Varun Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16860–16870, 2025. 9

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *International Conference on Computer Vision (ICCV)*, October 2025. 4

HyunJun Jung, Weihang Li, Shun-Cheng Wu, William Bittner, Nikolas Brasch, Jifei Song, Eduardo Pérez-Pellitero, Zhensong Zhang, Arthur Moreau, Nassir Navab, et al. Scrream: Scan, register, render and map: A framework for annotating accurate and dense 3d indoor scenes with a benchmark. *Advances in Neural Information Processing Systems*, 37:44164–44176, 2024. 6, 7, 16, 17

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

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision (ECCV)*, pp. 71–91. Springer, 2024. 3

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3D: Fast text-to-3D with sparse-view generation and large reconstruction model. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 3

Rui Li, Biao Zhang, Zhenyu Li, Federico Tombari, and Peter Wonka. Lari: Layered ray intersections for single-view 3d geometric reasoning. *arXiv preprint arXiv:2504.18424*, 2025a. 6, 7, 8, 9, 16

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025b. 2, 3, 5, 7, 9, 16, 17

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 5

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model, 2024. URL https://arxiv.org/abs/2408.16767. 2, 3

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:22226–22246, 2023a. 3

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 9298–9309, 2023b. 3

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9970–9980, 2024. 3

B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision (ECCV)*, 2020. 3

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012. 3

Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4209–4219, 2024. 2, 3

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10684–10695, 2022. 3

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9420–9429, 2024. 3

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=FUgrjq2pbB. 3

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2930–2937, 2013. 7, 8

Edgar Sucar, Zihang Lai, Eldar Insafutdinov, and Andrea Vedaldi. Dynamic point maps: A versatile representation for dynamic 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7295–7305, October 2025. 19

Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. Bolt3d: Generating 3d scenes in seconds. In *International Conference on Computer Vision (ICCV)*, October 2025. 2, 4

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision (ECCV)*, pp. 1–18. Springer, 2025a. 3

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025b. 1, 3

Tencent Hunyuan3D Team. Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation, 2024. 2, 3, 5

Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 2, 3, 5

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2, 3

Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:10021–10039, 2022. 2, 3

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30, 2017. 5

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision (ECCV)*, pp. 439–457. Springer, 2024. 3

Matthew Wallingford, Anand Bhattad, Aditya Kusupati, Vivek Ramanujan, Matt Deitke, Aniruddha Kembhavi, Roozbeh Mottaghi, Wei-Chiu Ma, and Ali Farhadi. From an image to a scene: Learning to imagine the world from a million 360° videos. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:17743–17760, 2024. 3

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 5294–5306, 2025a. 1, 2, 3, 4, 6, 7, 8, 16

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 52–67, 2018. 3

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025b. 1, 3, 6, 7, 8

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5261–5271, 2025c. 7

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024a. 1, 3, 4, 7

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b. 3

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 21551–21561, 2024. 3

Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 4

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025a. 2, 3, 17, 18

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21469–21480, 2025b. 7, 9

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025. 1, 3

Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. In *2024 International Conference on 3D Vision (3DV)*, pp. 664–674. IEEE, 2024. 3

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023. 7

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025. 3

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3

Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023. 2, 3, 5, 9

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 (ECCV)*, pp. 1–19. Springer, 2024a. 3

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024b. 2, 3

Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. 4

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025b. 1, 3

Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in neural information processing systems (NeurIPS)*, 36: 73969–73982, 2023. 3

Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9720–9731, 2024. 3

Jensen Jinghao Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 3

A Appendix

A.1 More implementation details.

Model architectures. For the 3D point autoencoder (Stage 1), we follow the point encoder design from TripoSG Li et al. (2025b), which consists of one cross-attention layer and eight self-attention layers. The initial point queries are obtained by farthest point sampling from the input point cloud, while the learnable queries are randomly initialized tokens. We use 512 tokens with dimension 64 for the object-level model and 768 tokens with dimension 128 for the scene-level model. For the flow-matching decoder, we use a joint block with two cross-attention layers and one self-attention layer. The goal is to enable self-attention–like information exchange among queries while keeping computation manageable. Concretely, each block (i) aggregates information from noisy query points into the scene tokens via cross-attention, (ii) performs self-attention on the scene tokens (small M) to mix global context efficiently, and (iii) projects the updated scene tokens back to the queries with a second cross-attention.

For the image-to-latent transformer in Stage 2, we follow the architecture of VGGT (Wang et al., 2025a), which alternates between local (frame-level) and global attention. Due to computational constraints, we adopt a 16-layer variant instead of the full 24-layer VGGT, initializing from its pretrained weights. We also reuse VGGT's image tokenizers with frozen weights to obtain image tokens. The initial 3D scene tokens are treated as a *3D frame* and share the same local attention mechanism with the image tokens. For the 3D scene token, we copy the camera token from the first view to enable reconstruction in the camera coordinate of the first view.

Training. We train our model in two stages. In Stage 1, we aggregate per-view point clouds into a single input cloud and apply farthest point sampling on a randomly selected subset to supervise the flow-matching decoder. Farthest point sampling ensures that the target point cloud is distributed more evenly, reducing the influence of overlapping points in visible regions. Stage 1 is trained for 50 epochs. In Stage 2, we reuse the flow-matching decoder from Stage 1 and train it together with our image encoder, initialized with pretrained VGGT weights. The same flow-matching loss is used in both stages. For object and scene completion, target point clouds are sampled from complete reconstructions. To demonstrate compatibility with pixel-aligned formats, we also train a variant using RGB-D input, where target point clouds are sampled from point maps back-projected from depth. Stage 2 is trained for another 50 epochs.

Regarding computational cost, the Stage-1 point encoder is lightweight and requires no paired image–point cloud data, enabling efficient training on large-scale synthetic 3D datasets. In practice, Stage 1 takes about 40% less training time than Stage 2, and inference remains single-stage, feed-forward, and efficient regardless of the two-stage setup. Overall, the two-stage design adds small overhead while substantially improving stability, data flexibility, and reconstruction quality.

Evaluation. For object- and scene-level completion, we follow Li et al. (2025a) and sample 10k points for the object task and 100k for the scene task. However, correspondence-based point cloud alignment is not applicable due to our non–pixel-aligned reconstruction. Instead, we optimize a 3D translation and a global (1D) scale relative to the ground-truth point cloud using Adam to improve alignment. We do not optimize rotation, as our reconstruction is expressed in the first-view coordinate frame.

A.2 More ablation study.

Reconstruction at any resolution. Since our approach models the point distribution rather than a per-pixel point map, it naturally supports resolution-agnostic generation by adjusting the number of noisy queries at inference. Figure 9 presents results with varying query counts for the flow-matching decoder, demonstrating that our method consistently produces point clouds at different resolutions with reliable reconstruction quality.

A.3 More visualizations.

We show more visualization results for scene-level completion on SCRREAM (Jung et al., 2024) dataset, as shonw in Figure 10. We also include density evaluation on NRGBD (Azinović et al.,

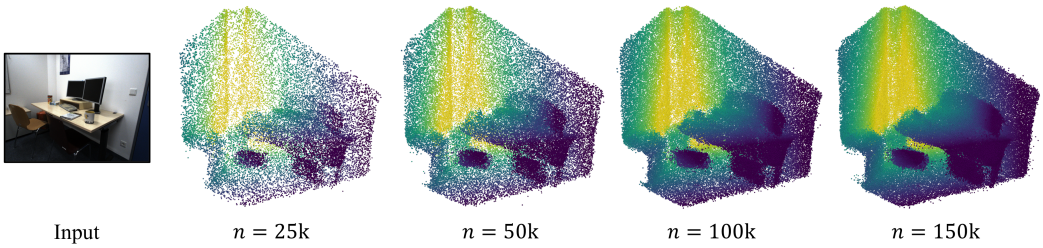

Figure 9: **Visualization of point cloud generation at different resolutions.** Our non–pixel-aligned formulation allows inference at arbitrary resolutions.

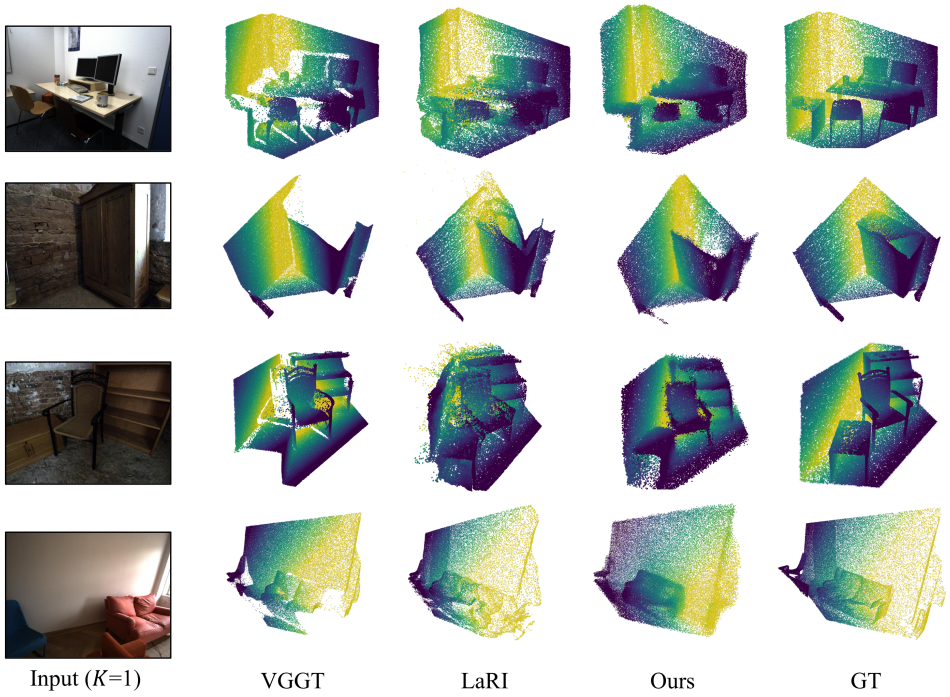

Figure 10: **Qualitative results for scene completion on SCRREAM Jung et al. (2024).** Our method shows better scene completion results compared to other baselines.

2022) in Figure 11. While trained with $K = 2$ views only, our method generalize to multiple image views ($K = 4$) and provides more evenly distributed point cloud .

A.4 Uncertainty comparison with Latent 3D Generation

Our method is specifically designed to reduce the uncertainty typically observed in latent diffusion-based 3D generation approaches such as TRELLIS (Xiang et al., 2025a) and TripoSG (Li et al., 2025b). These methods perform generation in a high-dimensional latent space, which often leads to hallucinated geometry, shape deviations, and inconsistencies across viewpoints—particularly when multiple input images are involved (see Figure 12). As a result, they struggle to maintain strong pixel-scene and cross-view alignment.

A.5 Performance on Outdoor Scenes

To validate the robustness and generalization capability of our framework, we further evaluate NOVA3R using the outdoor dataset Virtual KITTI 2 (Cabon et al., 2020). We finetune our model on Virtual KITTI 2 to better adapt to large-scale outdoor environments. To construct pseudo ground

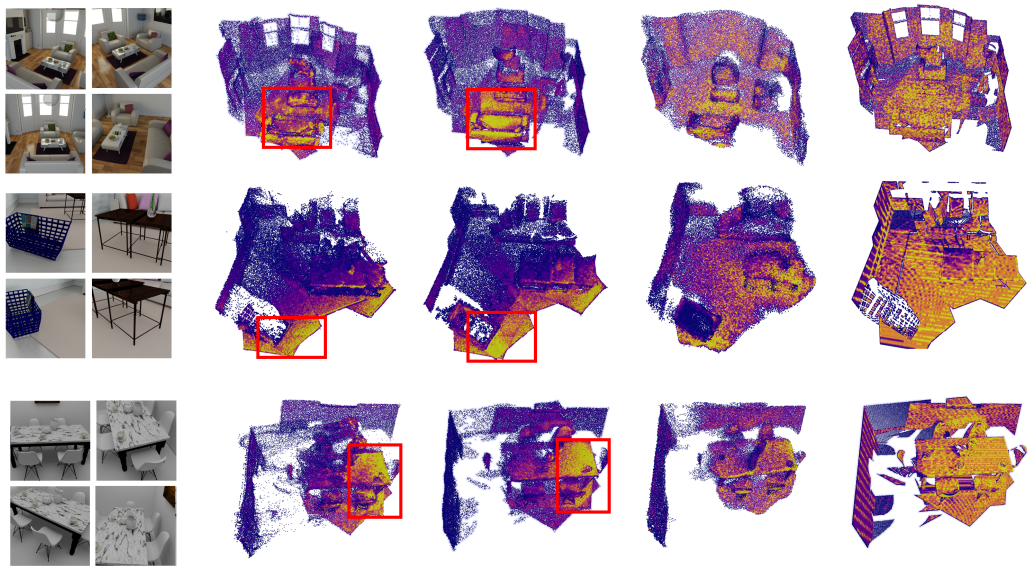

Figure 11: **Qualitative results for density evaluation on NRGBD** (K=4) **(Azinović et al., 2022).** Yellow regions denote higher density, and purple regions denote lower density. Our method provides more evenly-distributed point cloud (colored by density).

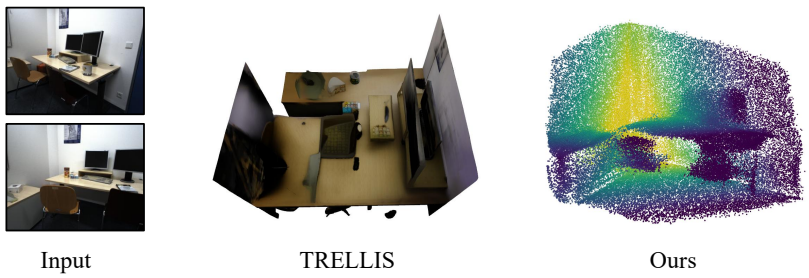

Figure 12: **Qualitative comparison for generation uncertainty on SCRREAM.** Our method produces more geometrically faithful reconstructions compared to TRELLIS (Xiang et al., 2025a).

truth, for each input frame we collect neighboring frames within [-4,8] timesteps and additional views from $\pm 15°$ and $\pm 30°$ viewpoints. Using depth maps and camera parameters, we project them into per-frame point clouds, transform them to world coordinates, and retain only points within the target view's frustum. As shown in Figure 13, NOVA3R performs well on outdoor scenes, further demonstrating its ability to handle both indoor and outdoor scenarios.

A.6 Discussion

Large-scale Scenes. Modeling large-scale scenes with many input images is a major computational bottleneck for existing learning-based 3D reconstruction methods, particularly for pixel-aligned approaches like VGGT, which must handle duplicated points across multiple views. In contrast, our point-wise decoding uses fewer tokens to represent the scene, making it inherently more scalable. However, the number of points needed varies across scenes of different scales, requiring adaptive point selection strategies, such as using sparse COLMAP point maps to guide point count. Furthermore, when processing a large number of images that cannot fit in GPU memory simultaneously, an online image encoder (e.g., CUT3R-style) with a fixed memory buffer can be used to enable incremental processing without increasing memory usage.

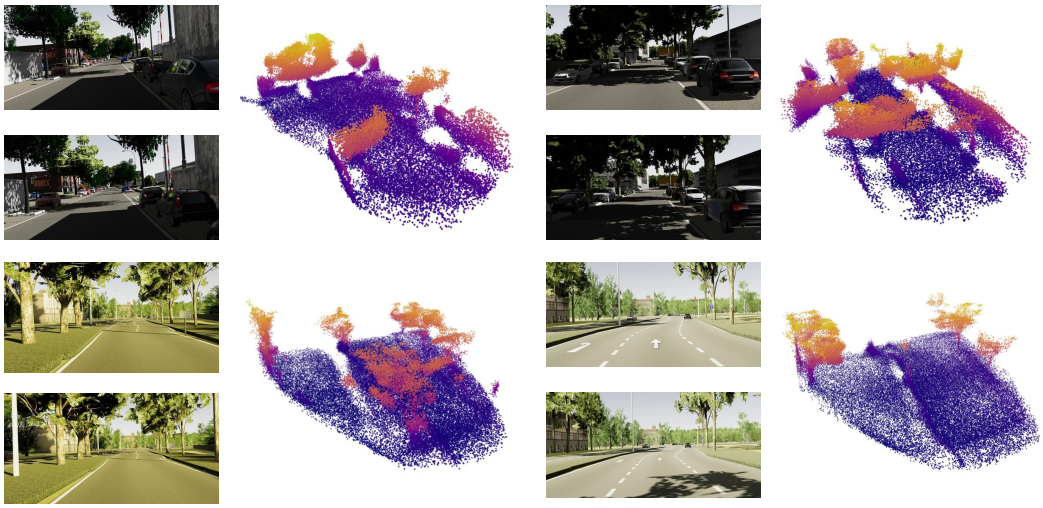

Figure 13: **Qualitative results for outdoor scenes reconstruction on Virtual KITTI 2.** Our method is also applicable to outdoor scene reconstruction (colored by Y axis).

Dynamic Scenes. Our paradigm is inherently extensible to dynamic scenes, either by adding a branch to predict target time point maps (Sucar et al., 2025; Feng et al., 2025) or by extending the 3D latent autoencoder to a time-conditioned 4D latent representation. Such a representation can potentially model the entire 4D scene more efficiently by capturing both complete geometry and temporal evolution across the whole sequence, rather than relying on per-frame reconstruction. However, this requires carefully designed training data and loss functions to address temporal coherence and motion ambiguity, which is beyond the scope of this work. In this paper, we focus on static scene-level completion from multi-view inputs, which remains a challenging and under-explored problem.