VolFill: Single-View Amodal 3D Scene Reconstruction with Volumetric Flow Matching

Tuan Duc Ngo[1] **Chuang Gan**[1] **Evangelos Kalogerakis**[1,2]
[1]UMass Amherst [2]Technical University of Crete
github.com/VolFill

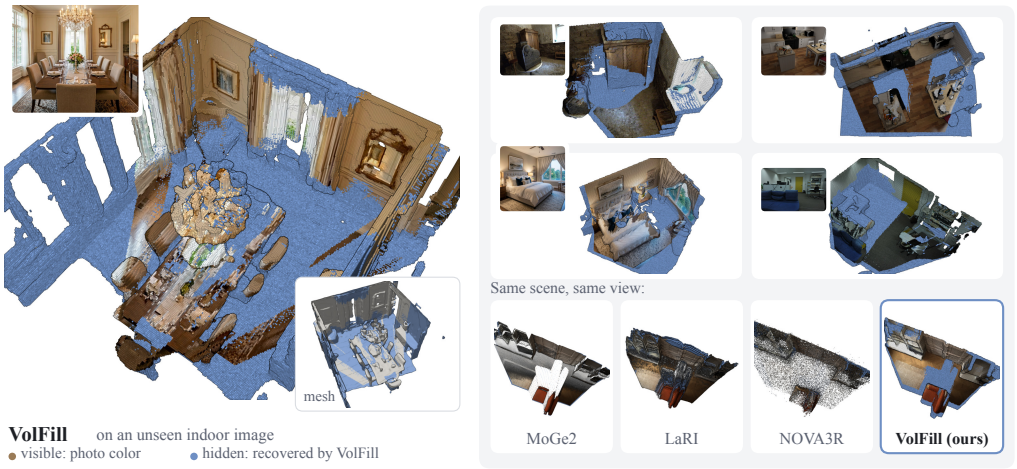


Figure 1: **VolFill** recovers the *complete* 3D scene, including surfaces hidden from the camera, from a single image. Geometry seen in the input view keeps the photo's color; hidden geometry recovered by VolFill is shown in **blue**. *Left:* an unseen indoor image, as a point cloud and as a mesh extracted directly from the predicted TUDF grid. *Top right:* more qualitative samples. *Bottom right:* on the same scene and view, pixel-aligned MoGe2 leaves holes behind occluders, LaRI produces layered artifacts, and NOVA3R yields noisy, disintegrated points, whereas VolFill produces sharp and complete geometry.

Abstract

Reconstructing the complete geometry of a scene from a single RGB image remains challenging—especially when inferring hidden structures where visual evidence is incomplete. We introduce VolFill, a generative framework that predicts the 3D structure of the complete scene rather than relying on traditional pixel-aligned regression. Our method utilizes a hybrid 3D VAE to compress sparse *truncated unsigned distance function* grids into a compact latent space, paired with a latent Diffusion Transformer that denoises this representation to recover the complete scene. We condition the generation on a geometry foundation model, leveraging rich spatial priors for robust reasoning. Unlike existing methods limited by per-ray constraints or unstructured point-cloud queries, VolFill provides a structured representation that supports direct surface extraction and occupancy queries at scale. Extensive experiments on the SCRREAM and NRGB-D datasets demonstrate that our approach significantly outperforms current amodal baselines, providing a robust foundation for holistic spatial understanding.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

This paper addresses amodal 3D scene reconstruction, seeking to recover the complete geometry, including observed and occluded structures, from a single RGB image. Beyond being inherently ill-posed, amodal reconstruction must infer hidden structures where geometric evidence is partially absent. This requires synthesizing under-determined regions that remain physically plausible and consistent with the observed scene. Such spatial awareness is critical for practical applications, as navigating or interacting with an environment requires a comprehensive understanding of the scene that extends well beyond the immediate line of sight.

Recent feed-forward methods [5, 6, 41, 61, 83, 84, 85, 86, 87, 93, 95] can recover accurate 3D structure from images in seconds, but all share a hard constraint: *pixel alignment*. Every predicted point lies on a source camera ray, limiting reconstruction strictly to visible surfaces and producing duplicated geometry in overlapping views. Optimization-based scene representations [42, 53, 56, 60, 75] reconstruct complete geometry but require dense captures and per-scene optimization, which can take hours to compute. Large reconstruction models [34, 105] and generative 3D models [46, 90, 91, 97] demonstrate impressive 3D reconstruction quality from sparse inputs, but are predominantly optimized for object-centric scenarios, limiting their applicability to complex scenes with unbounded extents and challenging amodal structures.

At the scene level, LaRI [45] proposes a *layered ray intersection* representation, regressing multiple ordered surface intersections per camera ray. While this enables amodal recovery within a pixel-aligned framework, it remains a partial solution; the fixed layer count under-represents densely occluded structures while wasting capacity in open regions. Furthermore, it frequently yields layered artifacts (the "multiple wall" effect) rather than accurately recovering the actual unobserved surfaces. Alternatively, NOVA3R [13] employs flow matching to directly generate *amodal 3D point clouds*, but often yields noisy and disintegrated outputs (Fig. 1).

To address these limitations, we propose VolFill, a latent diffusion framework that predicts a complete Truncated Unsigned Distance Function (TUDF) voxel grid from a single image. A TUDF encodes the distance to the nearest scene surface, including occluded regions, as a continuous scalar field, enabling direct surface extraction via isosurfacing [27] without requiring post-processing reconstruction of point clouds. Unlike point clouds, it scales well with scene complexity; unlike layered ray representations, it places no constraint on the number of recoverable occluded layers. VolFill consists of a hybrid 3D VAE that encodes the sparse TUDF grid into a compact dense latent space, and a Diffusion Transformer (DiT) trained with flow matching. To make amodal reasoning robust despite limited 3D training data, we introduce a dual-conditioning strategy that combines high-level image tokens with explicit visible geometry from a frozen MoGe2 [85] model, grounding the generation in the observed scene while leveraging the strong geometric priors of this foundation model.

Trained on 3D-FRONT [19] and ScanNet++ [98], VolFill achieves state-of-the-art amodal reconstruction on the SCRREAM [37] and NRGB-D [2] benchmarks, synthesizing high-fidelity amodal geometry with significantly greater sharpness and structural accuracy than existing amodal methods (Fig. 1). In summary, our main contributions are:

- **VolFill**, a generative framework that utilizes volumetric flow matching to recover complete scene-level geometry from single-view images.
- **A hybrid 3D VAE** enabling efficient spatial compression of high-resolution TUDF grids to a compact latent space, facilitating high-fidelity reconstruction of complex amodal structures.
- **A dual-conditioning strategy** leveraging geometry foundation models to integrate high-level image tokens with explicit visible geometry.

2 Related Work

Volumetric 3D representations provide a flexible foundation for 3D vision, spanning applications from scene understanding to generative modeling. To mitigate the cubic complexity of dense grids, sparse convolutional engines [15, 25, 77] have become the standard for efficiently processing geometry along active surfaces. In scene-level reconstruction, "lifting" paradigms map 2D features into these volumes through depth-guided projection [76], ray sampling [8], or transformer-based voxel queries [47, 88]. Within the generative domain, early models applied diffusion directly to raw

voxels [57], whereas contemporary frameworks leverage hierarchical latent diffusion [14, 67] or rectified flow on sparse latents [91] to achieve high-fidelity structural synthesis. Building on these advances, our approach adopts a TUDF representation within a structured 3D grid, using sparse convolutions to capture complex scene geometry while maintaining high computational efficiency.

Pixel-aligned single-view 3D reconstruction recovers *visible* surface geometry, evolving from early handcrafted features [33, 38, 70, 71] to deep architectures [3, 17, 20, 43, 81]. Progress in scaling has yielded models with remarkable zero-shot generalization [4, 5, 28, 35, 64, 66, 95, 96, 99]. Recently, methods have further refined geometric accuracy by jointly predicting intrinsics [5, 64], regressing dense 3D pointmaps [84, 85], or utilizing diffusion priors [65, 69] for high-fidelity, arbitrary-resolution depth synthesis [21, 23, 30, 40, 63, 93, 100]. Nonetheless, these approaches remain fundamentally confined to 2.5D visible surfaces. Their inability to infer occluded regions or amodal structure directly motivates our shift from surface-level estimation to full-scene reconstruction.

Geometry foundation models as priors. Recent geometry foundation models (GFMs) [18, 48, 83, 84, 85, 86] have demonstrated strong performance in 3D reconstruction and serve as robust feature extractors for downstream tasks requiring explicit spatial priors [10, 58, 89, 93]. Current approaches typically adopt one of two strategies: (1) using explicit geometric outputs (e.g., depth or point clouds) as structural anchors to guide synthesis [9, 12, 101, 102], or (2) injecting latent foundation features into transformer blocks to maintain consistency [36, 79, 89, 92]. Our approach unifies these paradigms via a dual-conditioning strategy. By leveraging both explicit visible geometry for structural grounding and foundation latent tokens for high-level context, we enable robust and physically plausible amodal reconstruction that transfers to held-out real indoor benchmarks.

Amodal 3D reconstruction aims to recover complete geometry from partial visual observations. Object-centric strategies often utilize two-stage pipelines—generating novel views via multi-view diffusion [50, 52, 54, 74] before reconstructing geometry [34, 44, 78, 94, 108]—or synthesize 3D representations directly [11, 46, 59, 91, 97], yet remain limited to isolated assets. For scenes, recent methods leverage camera trajectories for view synthesis [22, 29, 68, 82, 102, 106] but typically lack direct 3D outputs. Most related are LaRI [45], using layered ray intersections, and NOVA3R [13], which generates point clouds from scene tokens. **Concurrent with our work**, World Tracing [104] generates layered pixel-aligned geometry beyond the visible surface, while Surflo [26] and GenRecon [72] reconstruct scenes from multiple views. In contrast, our structured volumetric representation enables more complete and accurate amodal reconstruction, recovering geometry with smoother surfaces and sharper structural detail than unstructured point-based approaches.

3 Method

3.1 Problem Formulation

We address the task of amodal scene reconstruction from a single RGB image. Unlike traditional pixel-aligned methods [5, 84, 85, 86, 95] that recover only visible surfaces (e.g., *depth maps* or *pointmaps*), our goal is to predict the complete 3D geometry within the observed view frustum. This includes recovering amodal geometry, surfaces occluded by foreground objects, while maintaining the structural consistency of the scene-level environment.

Input and output. Given a single input image $I \in \mathbb{R}^{H\times W\times 3}$, our model predicts a complete volumetric grid $V \in \mathbb{R}^{N\times N\times N}$ at a resolution of $N = 256$. We represent the scene geometry using a *Truncated Unsigned Distance Function*. Let $\mathcal{S} \subset \mathbb{R}^3$ denote the set of all physical surfaces in the scene contained within the view frustum. For each voxel center p within our grid, the value $V(p)$ is defined as the distance to the nearest surface in $\mathcal{S}$, truncated to a maximum value τ:

$$V(p) = \min(\text{dist}(p, \mathcal{S}), \tau) \tag{1}$$

We employ an unsigned formulation because indoor scenes are frequently non-watertight. A signed distance function requires a globally consistent inside/outside orientation—well-defined only for closed, watertight surfaces. For the common open geometries (single-sided walls, floors, furniture), no such orientation exists, making the sign geometrically meaningless beyond the immediate voxel neighborhood. The unsigned formulation avoids this ambiguity, as distance-to-nearest-surface is well-defined for any geometry regardless of topology.

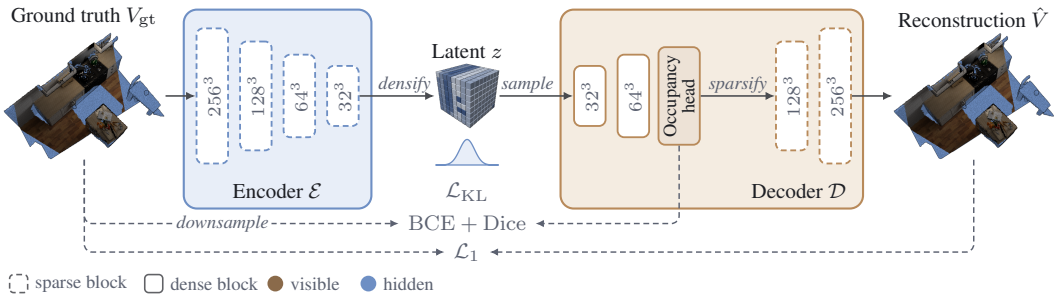


Figure 2: **3D VAE architecture.** The encoder compresses high-resolution sparse TUDF grids into a regularized dense latent via sparse convolutions. The decoder upsamples through dense layers, applies occupancy-guided sparsification, then restores the full-resolution TUDF via sparse convolutions.

Amodal geometry voxelization. The amodal surfaces $\mathcal{S}$ can be obtained either directly from ground-truth dataset 3D meshes (if available) or by fusing depth from multiple surrounding views and transforming it into the target camera coordinate frame. To translate these amodal surfaces into our discrete target grid V, we employ a dynamic, frustum-aligned discretization strategy:

- *Spatial scope:* To establish the precise boundaries of our volumetric grid, we compute an axis-aligned bounding box $\mathcal{B}$ based on the *visible* geometry $P_{\text{vis}} \subset \mathcal{S}$. Any amodal surfaces in $\mathcal{S}$ extending beyond $\mathcal{B}$ are discarded, focusing the representation on the immediate captured scene.
- *Dynamic scaling and TUDF computation:* Rather than using fixed spatial extents [8, 47, 88], we calculate a dynamic, isotropic voxel size derived from the maximum dimension of $\mathcal{B}$ and our target resolution N. We then discretize this bounded space and compute the truncated distance from each voxel center p to the cropped amodal surfaces, yielding the final TUDF grid V.

Downstream geometric extractions. The TUDF grid V serves as a flexible geometric proxy supporting diverse downstream formats. Surfaces can be efficiently extracted as point clouds by thresholding voxels with distance values below a small threshold $V(p) < \epsilon_{\text{tudf}}$, or converted into meshes via MeshUDF [27], producing clean, topologically consistent reconstructions.

3.2 Amodal 3D Reconstruction via a Generative Framework

We formulate amodal scene reconstruction as estimating the conditional distribution $P(V|I)$ rather than a deterministic point estimate, overcoming a critical limitation of regression-based pipelines [5, 84]. Standard regression losses (e.g., L_2) predict the conditional expectation $\mathbb{E}[V|I]$, averaging over plausible geometries and producing over-smoothed reconstructions in occluded regions where evidence is absent. In contrast, our generative framework samples from $P(V|I)$, producing sharp and physically plausible completions by following the learned manifold of real-world scene structures.

3.3 Hybrid 3D VAE

To bridge the gap between high-resolution TUDF grids and the manageable latent space required by our generative model, we propose a hybrid sparse-dense 3D VAE (Fig. 2). This architecture exploits the inherent sparsity of scene geometry, where typically only 3–5% of voxels are surface-adjacent, while providing a fixed-size dense latent representation suitable for denoising via a latent transformer.

The encoder $\mathcal{E}$ is designed to process natively sparse TUDF inputs with high computational efficiency. We convert the input grids into sparse tensors, retaining only active voxels within the truncation band τ, and apply a sequence of sparse 3D convolutions [15, 16, 77] with strided downsampling to reduce the spatial resolution by $16\times$. At the bottleneck, these features are densified into a regular 4D tensor $z \in \mathbb{R}^{16\times16\times16\times C}$; by densifying only at this highly compressed scale, we maintain a manageable memory footprint while ensuring compatibility with dense generative architectures. Finally, a standard KL-divergence bottleneck regularizes the latent space to facilitate stable sampling [69].

The decoder $\mathcal{D}$ faces the challenge of recovering a 256^3 grid from a dense 16^3 latent without prior knowledge of the target scene's sparsity pattern. We address this through a hybrid dense-to-sparse decoding schedule:

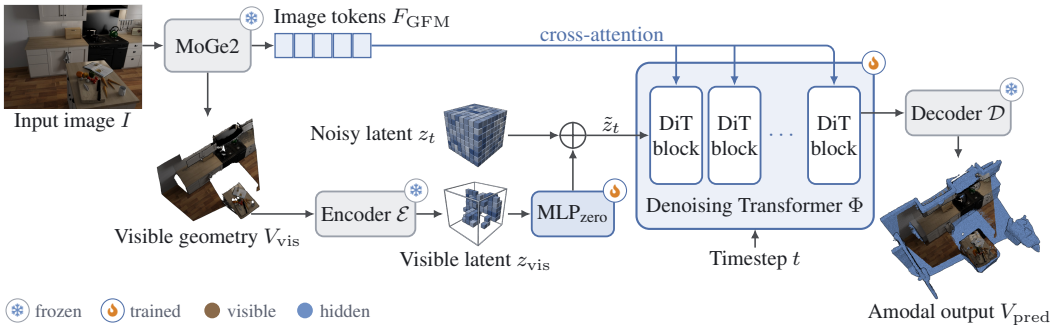


Figure 3: **Latent DiT architecture.** It operates in the compressed VAE latent space using a denoising transformer backbone with a flow-matching objective. We leverage a **dual-conditioning** strategy, integrating high-level image tokens and explicit visible geometry, to guide the generative process and synthesize sharp, scene-consistent amodal structures.

- *Dense upsampling* ($16^3 \rightarrow 64^3$): We first employ standard transposed 3D convolutions to upsample the latent to a 64^3 dense feature map, where the total voxel count ($\sim$ 262K) remains tractable for dense computation.
- *Structure prediction*: At the 64^3 stage, a binary occupancy head predicts a mask $\hat{O}$, indicating surface-adjacent voxels. This mask is used to sparsify the dense feature map, discarding empty regions and focusing subsequent computation solely on relevant geometric structures. We supervise this with a ground-truth occupancy mask O_{gt} derived from the TUDF grid.
- *Sparse upsampling* ($64^3 \rightarrow 256^3$): The sparse features are processed by sparse 3D convolutions to reach the final 256^3 resolution, where memory cost scales linearly with surface area rather than cubically with volume, enabling high-resolution TUDF reconstruction at surface-adjacent voxels. The resulting sparse predictions are finally scattered back into a dense grid to obtain $\hat{V}$.

The VAE is trained end-to-end on our collected TUDF datasets using a composite loss function:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_1(\hat{V}, V_{\text{gt}}) + \lambda_{\text{bce}}\text{BCE}(\hat{O}, O_{\text{gt}}) + \lambda_{\text{dice}}\text{Dice}(\hat{O}, O_{\text{gt}}) + \lambda_{\text{kl}}\mathcal{L}_{\text{KL}} \tag{2}$$

where $\mathcal{L}_1$ ensures precise distance reconstruction in active voxels, BCE and Dice losses supervise binary occupancy in the dense-to-sparse transition, and $\mathcal{L}_{\text{KL}}$ regularizes the latent distribution. This hybrid design facilitates aggressive $16\times$ spatial compression while preserving the fine-grained details essential for high-fidelity amodal reconstruction.

3.4 Geometry-Conditioned Flow Matching

We implement a latent Diffusion Transformer Φ that learns the conditional velocity field to transport noise toward the distribution of scene TUDF latents. Starting from $z_0 \sim \mathcal{N}(0, \mathbf{I})$, we iteratively integrate the learned flow to obtain a clean latent z_1, which is decoded by $\mathcal{D}$ to produce the final TUDF prediction V_{pred}. To guide this process, we propose a dual-conditioning strategy that supplements global semantic-geometric context with local structural anchors. This architecture is shown in Fig. 3.

Global geometric priors. To inherit a robust, zero-shot understanding of spatial layouts, we condition the DiT on frozen features (F_{GFM}) extracted from a geometry foundation model. We utilize MoGe2 [85] as our backbone for its strong geometric estimation performance. Following standard practice, these features are processed via layer normalization and a linear projection before being injected into the cross-attention layers of each DiT block as keys and values (details in Appendix A.2).

Visible-latent inpainting. While global features provide context, we explicitly anchor the generative process to the observed scene by integrating known visible geometry directly into the latent space. Conceptually motivated by unprojection-based synthesis [12, 102], this strategy provides a partial guide for the model to complete occluded regions. Concretely, we obtain the visible pointmap P_{vis} from the GFM and convert it into a TUDF grid. This is then encoded into a visible latent z_{vis} via the frozen VAE encoder $\mathcal{E}$, and fused with the noisy latent as:

$$\tilde{z}_t = z_t + \text{MLP}_{\text{zero}}(z_{\text{vis}}) \tag{3}$$

Table 1: **3D reconstruction results** on the SCRREAM dataset. We mark best and second-best.

Method	Visible			Occluded			Complete		
	CD↓	$APD_{0.1}$ ↑	$APD_{0.02}$ ↑	CD↓	$APD_{0.1}$ ↑	$APD_{0.02}$ ↑	CD↓	$FS_{0.1}$ ↑	$FS_{0.02}$ ↑
TripoSG† [46]	13.75	44.74	10.08	10.73	55.34	13.37	11.10	53.56	13.30
TRELLIS† [91]	11.63	54.05	15.00	9.65	60.70	17.52	10.72	56.69	15.59
DUSt3R [86]	4.46	92.13	30.81	23.21	38.43	8.40	10.85	69.90	24.17
VGGT [83]	3.04	96.98	45.62	17.40	44.60	10.04	6.77	80.45	36.77
DA3 [48]	2.37	98.98	54.34	16.16	48.31	10.76	6.16	82.86	43.10
DepthPro [5]	3.85	95.10	33.30	15.31	48.66	9.71	6.90	80.36	28.44
MoGe2 [85]	2.28	99.38	57.42	15.29	50.09	10.30	5.74	83.96	45.84
LaRI [45]	3.43	95.48	41.61	5.23	85.25	29.31	5.19	85.43	30.85
NOVA3R [13]	3.20	96.77	43.75	3.56	94.86	41.21	3.43	95.26	45.12
VolFill (ours)	2.84	96.10	56.73	3.46	92.70	55.53	3.03	95.03	54.83

Note: All reported numbers are scaled by a factor of 10^2. †: methods trained on object-centric datasets.

where MLP_{zero} is a zero-initialized projection layer. This provides the model with an explicit structural prior over observed regions to guide completion of occluded geometry.

Flow matching objective. We train our model using a rectified flow objective [1, 49, 51]. Given a ground-truth TUDF latent z_1 and Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, we define a linear interpolation path $z_t = (1-t)\epsilon + tz_1$. Our training objective is to minimize the expected mean squared error between the predicted velocity u_Φ and the constant velocity target $u^* = z_1 - \epsilon$:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{t,\epsilon,z_1}\left[\, \|u_\Phi(\tilde{z}_t, t, F_{\text{GFM}}) - u^*\|^2 \right] \tag{4}$$

During inference, we integrate the learned ODE from $t = 0$ to $t = 1$ via Euler steps, then decode the resulting latent through the VAE decoder $\mathcal{D}$ to produce the final TUDF grid V_{pred}.

3.5 Model Architecture

The 3D VAE employs a symmetric encoder-decoder architecture, where each stage consists of two ResNet-style blocks and doubles the channel dimension at each resolution step. The primary distinction lies in the convolution implementation: while the encoder uses sparse 3D convolutions across all stages for efficiency, the decoder utilizes standard dense convolutions at lower resolutions ($16^3 \rightarrow 64^3$) before transitioning to sparse convolutions for the remaining higher stages ($64^3 \rightarrow 256^3$). Finally, the binary occupancy and TUDF regression heads are both implemented as 2-layer MLPs.

The latent DiT features 12 transformer blocks with a 768-dimensional hidden state and 12-head self-attention [80]. Each block integrates image tokens F_{GFM} through cross-attention, utilizes AdaLN [62] for timestep conditioning, and employs a GELU feed-forward network [31] with $4\times$ expansion. To stabilize flow-matching training, we apply QK-normalization [32] via RMSNorm [103] before the attention layers. The detailed architecture of the DiT block is provided in Appendix A.2.

4 Experiments

4.1 Implementation Details

Training datasets. We train VolFill on a combination of 3D-FRONT [19] and ScanNet++ [98], following the data splits established by LaRI [45] and NOVA3R [13]. For the synthetic 3D-FRONT dataset, we utilize 18k room-level scenes to construct 96k image-TUDF pairs. For real-world ScanNet++ data, we aggregate multi-view depth maps and back-project them into a unified point cloud, followed by a filtering and voxelization pipeline to obtain amodal TUDF grids for 46k samples. All TUDF grids use $N = 256$ and truncation $\tau = 3$ voxels within a frustum-aligned box estimated from the visible geometry with robust percentile clipping and 10% padding.

Table 2: **3D reconstruction results** on the NRGB-D dataset ($\times 10^2$).

Method	Visible			Occluded			Complete		
	CD↓	$APD_{0.1}$ ↑	$APD_{0.05}$ ↑	CD↓	$APD_{0.1}$ ↑	$APD_{0.05}$ ↑	CD↓	$FS_{0.1}$ ↑	$FS_{0.05}$ ↑
DUSt3R [86]	4.46	92.13	69.08	23.21	38.43	22.18	10.85	69.90	51.61
VGGT [83]	3.25	96.51	81.63	21.38	42.15	26.35	9.44	73.85	58.91
DA3 [48]	2.69	98.68	89.24	21.65	43.34	28.15	8.88	74.87	63.52
DepthPro [5]	4.38	92.56	67.41	22.19	41.69	24.64	10.03	71.30	51.60
MoGe2 [85]	2.62	98.46	89.55	21.12	43.31	28.43	9.31	75.35	61.93
LaRI [45]	4.09	93.77	73.51	13.04	64.53	45.82	9.79	68.37	43.47
NOVA3R [13]	5.05	89.63	64.44	8.31	80.50	58.55	8.19	73.99	49.92
VolFill (ours)	4.52	90.06	69.22	6.97	81.41	62.48	6.44	84.35	63.92

Training protocol. Training is conducted on two NVIDIA A6000/L40S GPUs in two stages: in Stage 1, the 3D VAE is trained for 20 epochs with a total batch size of 24, requiring 3 days to converge; in Stage 2, the latent transformer is trained for 100 epochs with a total batch size of 48, completing in 2.5 days. Both stages utilize the AdamW optimizer with an initial learning rate of 10^{-4}, a cosine scheduler, and mixed-precision training. More training details are provided in Appendix A.3.

Latent sampling and inference. During training, we sample the flow-matching timestep $t \in [0, 1]$ from a logit-normal distribution to prioritize training on higher-noise regimes following [91]. We apply classifier-free guidance (CFG) by dropping conditions with a probability of 0.1. At inference, we utilize an ODE solver with 50 steps and set the CFG scale to 3.0. On a single RTX 4090, inference takes 1.4 s per image and peaks at about 3.0 GB of GPU memory for generation and 3.5 GB end to end, including VAE decoding. Inference code and pretrained models are publicly available; training, evaluation, and data-processing code, together with the processed TUDF data and split files, will also be released.

4.2 Comparison with Prior Work

Datasets. We evaluate our model on two benchmarks. (1) SCRREAM [37] consists of 460 samples with complete, high-quality scanned meshes, providing reliable ground truth for both visible and occluded surfaces. (2) Neural RGB-D (NRGB-D) [2] is adapted for amodal evaluation by aggregating its depth maps into a global coordinate frame via camera trajectories and multi-view fusion to reconstruct occluded surfaces. To ensure the benchmark tests scene completion, we retain only frames with substantial hidden geometry using an objective rule: we crop the fused geometry to each frame's view frustum and keep frames whose amodal-enrichment ratio (occluded to visible voxels) is at least 0.2, yielding 137 samples from 9 scenes. We will release the curation script and the sample list.

Metrics. Following [13, 45], we evaluate reconstruction quality using Chamfer distance (CD↓) and F-score (FS_γ ↑) at thresholds $\gamma \in \{0.1, 0.02\}$ ($\{0.1, 0.05\}$ on NRGB-D). Since our non-pixel-aligned formulation lacks point-wise correspondences, we recover the optimal similarity transformation (scale, rotation, translation) by minimizing Chamfer distance via gradient descent prior to evaluation; this 7-DoF alignment and the evaluation protocol of NOVA3R [13] are applied identically to all methods. We assess performance across *visible*, *occluded*, and *complete* subsets. We employ a one-way Chamfer distance (measuring from ground truth to prediction) in the *visible* and *occluded* subsets to assess regional coverage without penalizing valid predicted geometry that falls outside the target subset, and replace the F-score with a threshold coverage score APD_γ ↑, measuring the percentage of ground-truth points successfully reconstructed within threshold γ. We further evaluate generative quality via Fréchet Point Cloud Distance (FPD↓) [55], using a pretrained Uni3D [107] to measure distributional similarity in a semantically aligned 3D feature space.

Baselines. We compare against four groups. *Object-level generative models* (TRELLIS [91], TripoSG [46]) show the effect of naively applying strong 3D object priors to full unmasked scenes. *Multi-view geometry* models (DUSt3R [86], VGGT [83], DepthAnything3 [48]) and *single-view pixel-aligned* estimators (MoGe2 [85], DepthPro [5]) establish visible-surface reference performance

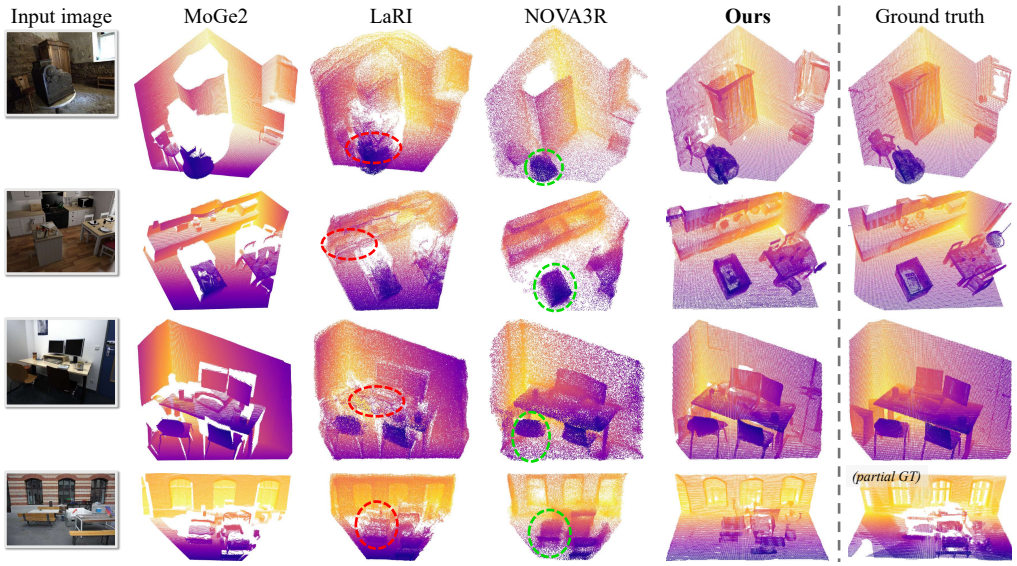


Figure 4: **Qualitative comparison.** VolFill synthesizes sharp, high-fidelity geometry, whereas LaRI produces layered artifacts (red circle) and holes, and NOVA3R yields noisy, unstructured point scatters (green circle).

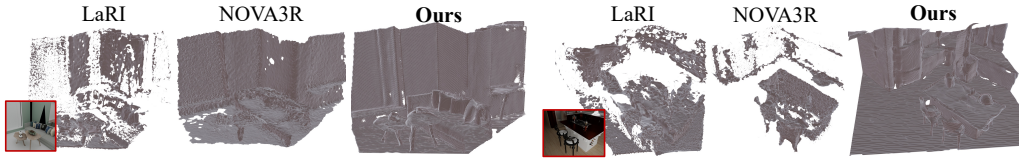


Figure 5: **Mesh reconstruction comparison.** LaRI and NOVA3R produce fragmented and noisy meshes due to their unstructured outputs, whereas VolFill directly extracts clean, topologically consistent surfaces from the structured TUDF grid.

but fundamentally cannot recover hidden geometry. Finally, *scene-level amodal reconstruction* methods (LaRI [45], NOVA3R [13]) serve as our primary competitors.

Results. Tables 1 and 2 report performance on SCRREAM and NRGB-D, respectively. While pixel-aligned estimators [48, 85] naturally dominate visible metrics, they inherently fail to reconstruct occluded regions. Among methods designed for amodal completion [13, 45], our approach establishes a new state of the art for complete scene reconstruction, achieving a significant performance leap under the stringent $\text{FS}_{0.02}$ metric on SCRREAM. On NRGB-D visible metrics, we maintain performance near that of LaRI [45], while achieving substantially stronger reconstruction across occluded and complete geometry—confirming that our volumetric formulation provides a far more robust foundation for full-scene understanding than existing point-cloud or ray-based baselines.

Is the representation the bottleneck? Our visible accuracy is slightly below that of MoGe2 (Table 1). To locate this gap, we pass ground-truth geometry through our representation on all 460 SCRREAM samples and score the recovered visible surface against the ground truth: converting it into our 256^3 TUDF grid and extracting the surface back yields a median visible $\text{APD}_{0.02}$ of 99.9, and additionally encoding and decoding it with our VAE also yields 99.9. As our full model reaches 66.3 on the same measure, neither the grid resolution nor the VAE limits accuracy; the gap arises in the generative step, which re-synthesizes the visible surface from soft conditioning rather than copying it.

Seed stability and diversity. All tables report a single sample per image. Over five sampling seeds on SCRREAM, results are stable (complete CD 3.11 ± 0.08, $\text{FS}_{0.02}$ 54.2 ± 0.5), while the variation between samples is about $1.3\times$ larger in occluded than in visible regions, indicating that VolFill produces diverse completions of ambiguous hidden geometry while keeping the observed surface consistent.

Table 3 presents our metric-scale evaluation, where we assess the reconstruction of amodal point clouds by applying a rigid transformation to align the predicted and ground-truth coordinates. While

Table 3: **Metric geometry evaluation.**

Method	CD↓	$FS_{0.02}$ ↑	FPD↓
LaRI	11.43	11.47	–
NOVA3R	6.56	22.23	16.50
VolFill (ours)	**3.88**	**44.90**	**4.86**

Table 4: **VAE decoder design ablation.**

Decoder	L1 Dist↓	IoU↑	Time (s)↓	Mem. (GB)↓
Dense-only	0.49	93.2	0.63	9.2
Sparse-only	0.48	93.0	0.09	2.5
Hybrid (ours)	**0.43**	**93.5**	**0.07**	**1.7**
w/ GT occupancy	0.44	93.5	0.07	1.7
w/ non-learned mask	1.79	74.0	0.08	1.8

Table 5: **Conditioning ablation.**

Config.	F_{GFM}	z_{vis}	CD↓	$FS_{0.02}$ ↑
Single	✗	*Add*	6.80	31.72
	MoGe2	✗	3.81	46.80
Dual	*MoGe2*	*Concat*	3.66	48.94
	MoGe2	*Add*	**3.03**	**54.83**

Table 6: **Foundation model ablation** (image-only conditioning).

Conditioning	CD↓	$FS_{0.02}$ ↑
DINOv2	5.46	35.98
VGGT	4.32	42.30
MoGe2	**3.81**	**46.80**

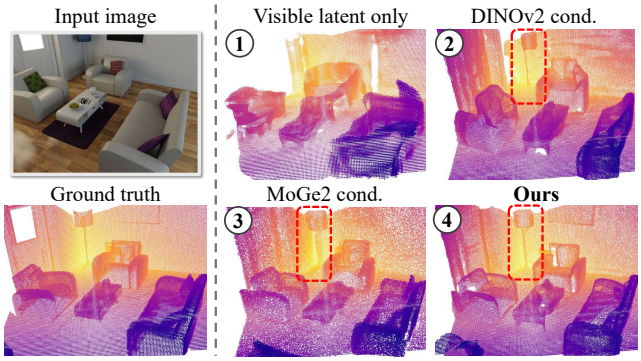


Figure 6: **Qualitative conditioning ablation.** ① Visible-only geometry fails to complete the scene; ②, ③ image-only tokens yield distorted results; ④ our dual-conditioning synthesizes sharp, high-fidelity amodal geometry.

our model leverages internal MoGe2 predictions to infer absolute scale, we provide baselines [13, 45] with a standalone MoGe2 model to lift their normalized outputs into metric space. Our approach significantly outperforms these methods, confirming its ability to reconstruct faithful metric-scale amodal geometry. Our superior FPD score demonstrates that the synthesized geometry is distributionally well-aligned with ground-truth scenes within a semantically meaningful feature space.

Fig. 4 presents qualitative comparisons of predicted point clouds. LaRI produces severe layered artifacts inherited from its ray-aligned formulation and leaves many occluded regions empty; NOVA3R recovers broader scene coverage but produces a noisy, unstructured point cloud. VolFill recovers more complete geometry with higher structural fidelity, especially in occluded regions.

Furthermore, the structured nature of our TUDF representation enables direct surface extraction [27], yielding smooth and topologically consistent meshes, as qualitatively evidenced in Fig. 5. In contrast, [13, 45] require Poisson reconstruction [39] to produce meshes, a slower post-processing step that nevertheless results in significant noise and disconnected regions wherever predicted density is sparse—further demonstrating that representing amodal scenes as structured distance functions is essential for recovering physically plausible surface geometry. See Appendix A.6 for more results.

4.3 Ablation Study

We evaluate our core design choices through ablation studies on the SCRREAM dataset to justify our architectural components and conditioning strategies for effective amodal scene reconstruction.

VAE decoder design is analyzed in Table 4. We evaluate reconstruction fidelity using the TUDF L1 distance within active cells and the occupancy Intersection over Union (IoU) from the thresholded TUDF. While a dense-only decoder is highly resource-intensive, a purely sparse variant reduces overhead. Both sparse variants share the same supervised occupancy head and differ only in the resolution at which it selects active voxels: filtering at the 16^3 bottleneck expands each active cell into a coarse block at 256^3, whereas our hybrid design upsamples densely to 64^3 before filtering, isolating surface voxels more tightly. It thus achieves the best reconstruction quality alongside the lowest latency and memory footprint, validating its efficiency for compressing sparse 3D structures into a compact latent space. The supervised occupancy head is both near-optimal and essential: replacing

its prediction with ground-truth occupancy leaves reconstruction unchanged (the head reaches 99.8% occupancy IoU), whereas a non-learned mask collapses it (L1 1.79, IoU 74.0).

Ablation on visual and geometric priors is shown in Table 5 and Fig. 6. Relying exclusively on the visible latent (①) fails to recover unobserved structures. While image-only conditioning (F_{GFM}) enables amodal reasoning, using DINOv2 (②) leads to distorted geometry, and replacing it with MoGe2 (③)—though more realistic—remains insufficient for recovering sharp, fine-grained details. Our full dual-conditioning strategy (④) achieves the highest fidelity, where additive fusion via zero-initialized MLPs (*Add*) outperforms concatenation (*Concat*), demonstrating that the synergy of global visual context and explicit geometric grounding is essential for accurate amodal reconstruction.

Ablation on different geometry foundation models is analyzed in Table 6. Despite its semantic richness, DINOv2 performs worst due to a lack of 3D spatial anchors. Geometry models like VGGT significantly improve results, while MoGe2 achieves the best performance. This suggests that structural consistency is a stronger driver for amodal reconstruction than general semantic reasoning.

5 Limitations and Conclusion

Limitations. Small holes can occasionally persist in heavily occluded regions, and since the scene is generated jointly, visible surfaces are slightly less sharp than those of MoGe2; reinstating the observed geometry (a training-free variant is evaluated in Appendix A.4) or adding explicit 3D supervision are natural remedies. Our training and evaluation focus on indoor scenes; outdoor or very large scenes would benefit from more diverse training data and spatial chunking of the grid. We also fix the grid at 256^3; since the decoder is largely sparse, its cost scales with the number of occupied voxels, so 512^3 or 1024^3 grids are feasible but require retraining the VAE and DiT, which we leave to future work. Finally, iterative sampling (1.4 s on an RTX 4090) remains slower than single-pass regression.

Conclusion. We introduced VolFill, a generative framework for amodal 3D scene reconstruction. A hybrid 3D VAE compresses sparse TUDF grids into a compact latent space, while a latent Diffusion Transformer denoises this representation to recover the complete scene geometry; frozen geometric priors from MoGe2 provide the necessary spatial context. This shifts reconstruction from deterministic pixel-aligned regression to a structured generative manifold, enabling physically plausible completions in occluded regions. VolFill produces sharper geometry and outperforms existing amodal point-cloud and ray-based methods.

Acknowledgments and Disclosure of Funding

Evangelos Kalogerakis has received funding from the European Research Council (ERC) under the Horizon research and innovation programme (Grant agreement No. 101124742).

References

[1] Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023.

[2] 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*, pages 6290–6301, 2022.

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021.

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[5] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*, 2025.

[6] Minh-Quan Viet Bui, Jaeho Moon, and Munchurl Kim. AirSplat: Alignment and rating for robust feed-forward 3D gaussian splatting. In *European Conference on Computer Vision (ECCV)*, pages 531–548, 2026.

[7] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision*, 2012.

[8] Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D semantic scene completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3991–4001, 2022.

[9] Anh-Quan Cao and Tuan-Hung Vu. Occany: Generalized unconstrained urban 3d occupancy. In *CVPR*, 2026.

[10] Jiahao Chang, Chongjie Ye, Yushuang Wu, Yuantao Chen, Yidan Zhang, Zhongjin Luo, Chenghong Li, Yihao Zhi, and Xiaoguang Han. ReconViaGen: Towards accurate multi-view 3D object reconstruction via generation. In *International Conference on Learning Representations (ICLR)*, 2026.

[11] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion NeRF: A unified approach to 3D generation and reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2416–2425, 2023.

[12] Kaihua Chen, Tarasha Khurana, and Deva Ramanan. Reconstruct, inpaint, test-time finetune: Dynamic novel-view synthesis from monocular videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[13] Weirong Chen, Chuanxia Zheng, Ganlin Zhang, Andrea Vedaldi, and Daniel Cremers. Nova3r: Non-pixel-aligned visual transformer for amodal 3d reconstruction. In *The Fourteenth International Conference on Learning Representations*, 2026.

[14] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G. Schwing, and Liang-Yan Gui. SDFusion: Multimodal 3D shape completion, reconstruction, and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4456–4465, 2023.

[15] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[16] Spconv Contributors. Spconv: Spatially sparse convolution library. `https://github.com/traveller59/spconv`, 2022.

[17] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[18] Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, and Chengfei Lyu. Dens3R: A foundation model for 3D geometry prediction. In *International Conference on Learning Representations (ICLR)*, 2026.

[19] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D furnished rooms with layouts and semantics. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10933–10942, 2021.

[20] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.

[21] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024.

[22] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: Create anything in 3D with multi-view diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 75468–75494, 2024.

[23] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan De Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 753–762. IEEE, 2025.

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32:1231 – 1237, 2013.

[25] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *ArXiv*, abs/1706.01307, 2017.

[26] Antoine Guédon, Shu Nakamura, Nicolas Dufour, Jiahui Lei, Ko Nishino, and Angjoo Kanazawa. Surflo: Consistent 3d surface flow model with global state. *arXiv preprint arXiv:2606.13644*, 2026.

[27] Benoît Guillard, Federico Stella, and Pascal Fua. MeshUDF: Fast and differentiable meshing of unsigned distance field networks. In *European Conference on Computer Vision (ECCV)*, pages 576–592. Springer, 2022.

[28] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rare**c**s Ambru**c**s, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9233–9243, 2023.

[29] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

[30] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations (ICLR)*, 2025.

[31] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

[32] Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253. Association for Computational Linguistics, 2020.

[33] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007.

[34] 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.

[35] 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*, 46(12):10579–10596, 2024.

[36] Wooseok Jang, Seonghu Jeon, Jisang Han, Jinhyeok Choi, Minkyung Kwon, Seungryong Kim, Saining Xie, and Sainan Liu. Repurposing geometric foundation models for multi-view diffusion. In *European Conference on Computer Vision (ECCV)*, pages 609–627. Springer, 2026.

[37] HyunJun Jung, Weihang Li, Shun-Cheng Wu, William Bittner, Nikolas Brasch, Jifei Song, Eduardo Pérez-Pellitero, Zhensong Zhang, Arthur Moreau, Nassir Navab, and Benjamin Busam. SCRREAM: SCan, register, REnder and map: A framework for annotating accurate and dense 3D indoor scenes with a benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 37, 2024.

[38] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth Transfer: Depth Extraction from Video Using Non-Parametric Sampling . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 36(11):2144–2158, November 2014.

[39] Michael M. Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Eurographics Symposium on Geometry Processing*, 2006.

[40] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9492–9502, 2024.

[41] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel López-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*, pages 499–509, 2026.

[42] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

[43] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[44] 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 *International Conference on Learning Representations (ICLR)*, 2024.

[45] Rui Li, Biao Zhang, Zhenyu Li, Federico Tombari, and Peter Wonka. LaRI: Layered ray intersections for single-view 3D geometric reasoning. In *International Conference on Machine Learning (ICML)*, 2026.

[46] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, and Yan-Pei Cao. TripoSG: High-fidelity 3D shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[47] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9087–9098, 2023.

[48] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.

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

[50] 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9264–9275, 2023.

[51] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

[52] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations (ICLR)*, 2024.

[53] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019.

[54] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3D: Single image to 3D using cross-domain diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9970–9980, 2024.

[55] Quan Meng, Yujin Chen, Lei Li, Matthias Nießner, and Angela Dai. Seen2Scene: Completing realistic 3D scenes with visibility-guided flow. In *European Conference on Computer Vision (ECCV)*, pages 386–404, 2026.

[56] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[57] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. DiffRF: Rendering-guided 3D radiance field diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4328–4338, 2023.

[58] Tuan Duc Ngo, Jiahui Huang, Seoung Wug Oh, Kevin Blackburn-Matzen, Evangelos Kalogerakis, Chuang Gan, and Joon-Young Lee. Dage: Dual-stream architecture for efficient and fine-grained geometry estimation. In *CVPR*, 2026.

[59] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *ArXiv*, abs/2212.08751, 2022.

[60] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3512, 2020.

[61] Jongmin Park, Minh-Quan Viet Bui, Juan Luis Gonzalez Bello, Jaeho Moon, Jihyong Oh, and Munchurl Kim. EcoSplat: Efficiency-controllable feed-forward 3D gaussian splatting from multi-view images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[62] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2023.

[63] Duc-Hai Pham, Tung Do, Phong Nguyen, Binh-Son Hua, Khoi Nguyen, and Rang Nguyen. SharpDepth: Sharpening metric depth predictions using diffusion distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17060–17069, 2025.

[64] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10106–10116, 2024.

[65] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

[66] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022.

[67] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. XCube: Large-scale 3D generative modeling using sparse voxel hierarchies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4209–4219, 2024.

[68] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6121–6132, 2025.

[69] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022.

[70] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems (NIPS)*, volume 18, 2005.

[71] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2009.

[72] Katharina Schmid, Nicolas von Lützow, Jozef Hladký, Angela Dai, and Matthias Nießner. Genrecon: Bridging generative priors for multi-view 3d scene reconstruction. *arXiv preprint arXiv:2605.23888*, 2026.

[73] 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*, pages 3260–3269, 2017.

[74] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *International Conference on Learning Representations (ICLR)*, 2024.

[75] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

[76] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 190–198, 2017.

[77] Haotian Tang, Zhijian Liu, Xiuyu Li, Yujun Lin, and Song Han. TorchSparse: Efficient Point Cloud Inference Engine. In *Conference on Machine Learning and Systems (MLSys)*, 2022.

[78] 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*, 2024.

[79] Tri Ton, Ji Woo Hong, and Chang D. Yoo. TARO: Timestep-adaptive representation alignment with onset-aware conditioning for synchronized video-to-audio synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14228–14237, 2025.

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[81] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2022–2030, 2018.

[82] Chaoyang Wang, Peiye Zhuang, Tuan Duc Ngo, Willi Menapace, Aliaksandr Siarohin, Michael Vasilkovsky, Ivan Skorokhodov, Sergey Tulyakov, Peter Wonka, and Hsin-Ying Lee. 4Real-Video: Learning generalizable photo-realistic 4D video diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17723–17732, 2025.

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

[84] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5261–5271, 2025.

[85] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 40377–40408, 2025.

[86] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024.

[87] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations (ICLR)*, 2026.

[88] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21672–21683, 2023.

[89] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3D representation for consistent world modeling. In *International Conference on Learning Representations (ICLR)*, 2026.

[90] Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3D generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14419–14429, 2026.

[91] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21469–21480, 2025.

[92] Mingyang Xie, Numair Khan, Tianfu Wang, Naina Dhingra, Seonghyeon Nam, Haitao Yang, Zhuo Hui, Christopher Metzler, Andrea Vedaldi, Hamed Pirsiavash, and Lei Luo. LaVR: Scene latent conditioned generative video trajectory re-rendering using large 4D reconstruction models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25450–25460, 2026.

[93] Gangwei Xu, Haotong Lin, Hongcheng Luo, Xianqi Wang, Jingfeng Yao, Lianghui Zhu, Yuechuan Pu, Cheng Chi, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Sida Peng, and Xin Yang. Pixel-perfect depth with semantics-prompted diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 193525–193549, 2025.

[94] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. GRM: Large gaussian reconstruction model for efficient 3D reconstruction and generation. In *European Conference on Computer Vision (ECCV)*, pages 1–20, 2024.

[95] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024.

[96] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[97] Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, Lifu Wang, Jing Xu, Zebin He, Zhuo Chen, Sicong Liu, Junta Wu, Yihang Lian, Shaoxiong Yang, Yuhong Liu, Yong Yang, Di Wang, Jie Jiang, and Chunchao Guo. Hunyuan3D 1.0: A unified framework for text-to-3D and image-to-3D generation. arXiv preprint arXiv:2411.02293, 2024.

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

[99] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9009–9019, 2023.

[100] Hao Yu, Haotong Lin, Jiawei Wang, Jiaxin Li, Yida Wang, Xueyang Zhang, Yue Wang, Xiaowei Zhou, Ruizhen Hu, and Sida Peng. InfiniDepth: Arbitrary-resolution and fine-grained depth estimation with neural implicit fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26920–26930, 2026.

[101] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 100–111, October 2025.

[102] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025.

[103] Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems 32*, Vancouver, Canada, 2019.

[104] Hao Zhang, Mohamed El Banani, Jen-Hao Cheng, Paul Zhang, Yi Hua, Ben Mildenhall, Christoph Lassner, Narendra Ahuja, and Gengshan Yang. World tracing: Generative pixel-aligned geometry beyond the visible. *arXiv preprint arXiv:2606.13652*, 2026.

[105] 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)*, pages 1–19, 2024.

[106] Jensen 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. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12405–12414, 2025.

[107] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring unified 3D representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024.

[108] Peiye Zhuang, Songfang Han, Chaoyang Wang, Aliaksandr Siarohin, Jiaxu Zou, Michael Vasilkovsky, Vladislav Shakhrai, Sergey Korolev, Sergey Tulyakov, and Hsin-Ying Lee. GTR: Improving large 3D reconstruction models through geometry and texture refinement. In *International Conference on Learning Representations (ICLR)*, 2025.

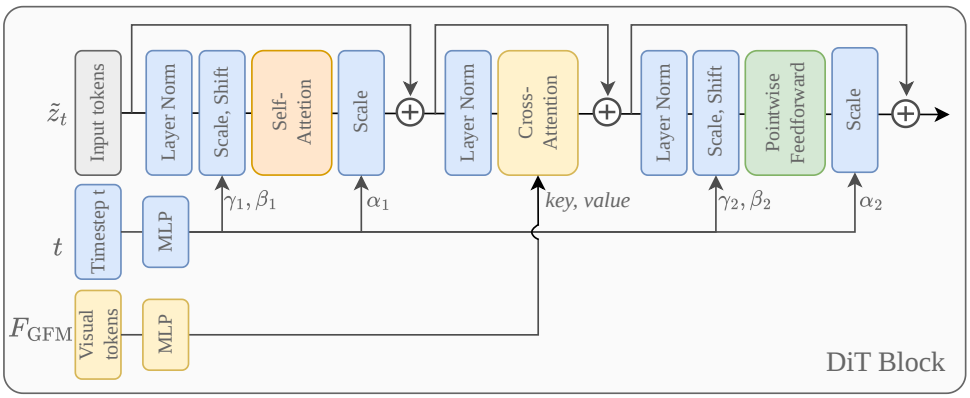


Figure 7: Illustration of the DiT block.

A Technical Appendices and Supplementary Material

A.1 Volumetric Representation Design Choice

We represent amodal 3D geometry as a Truncated Unsigned Distance Function (TUDF), where each voxel stores its distance to the nearest surface, clipped to a maximum of τ voxels. This choice is motivated by the limitations of the two alternatives below.

Binary occupancy marks each voxel as occupied or empty, providing only a single bit of information per voxel, while TUDF encodes the continuous distance to the nearest surface. This continuous formulation provides smooth gradients that guide the decoder toward surface boundaries, whereas binary losses offer no directional signal within the truncation band. Additionally, TUDF enables precise sub-voxel surface localization through isosurface extraction, while binary occupancy is restricted to discrete voxel centers.

Truncated Signed Distance Functions (TSDF) require a globally consistent inside/outside orientation, which is well-defined only for closed, watertight surfaces. Indoor scenes are frequently non-watertight (single-sided walls, floors, and furniture surfaces do not enclose a volume), so no such orientation exists. Consequently, TSDF provides no practical advantage over TUDF for open-geometry scenes. The unsigned formulation avoids this ambiguity entirely, as distance-to-nearest-surface remains well-defined regardless of scene topology.

A.2 Architecture Details

Fig. 7 describes the internal components of our DiT block, which processes latent tokens $\tilde{z}_t$ through sequential self-attention, cross-attention, and pointwise feed-forward stages. To inject temporal information, the timestep t is mapped via an MLP to generate adaptive layer normalization (AdaLN) parameters $\{\gamma_i, \beta_i, \alpha_i\}$ that modulate and scale the features. Visual priors are integrated by passing foundation model tokens F_{GFM} through an MLP to serve as key and value pairs for cross-attention.

A.3 Implementation Details

For the construction of the TUDF grids (Sec. 3.1), the truncation distance τ is set to 3 voxels. In the VAE architecture (Sec. 3.3), the bottleneck latent space ($16\times$ downsampling) utilizes a channel dimension of $C = 16$. To derive the low-resolution ground-truth binary occupancy mask from the high-resolution TUDF grid, we apply 3D max-pooling to accurately preserve the structural bounds of the geometry. Finally, the weighting coefficients for the composite VAE objective (Eq. 2) are empirically balanced and set to $\lambda_{\text{bce}} = 0.2$, $\lambda_{\text{dice}} = 0.2$, and $\lambda_{\text{kl}} = 10^{-6}$.

NRGB-D benchmark. As described in Sec. 4, we keep the NRGB-D frames whose fused geometry, cropped to the frame's view frustum, contains at least 0.2 occluded voxels per visible voxel. Table 7 lists the per-scene breakdown of the resulting 137 samples.

Table 7: **NRGB-D evaluation set**: samples per scene after the amodal-enrichment filter.

Scene	breakfast room	complete kitchen	green room	grey-white room	kitchen	morning apartment	staircase	thin geometry	white-room	**Total**
Samples	17	15	12	14	15	13	20	11	20	**137**

A.4 Training-Free Visible Reinstatement

Our visible surfaces are slightly less accurate than those of MoGe2 (Tables 1 and 2) because VolFill generates the visible and hidden geometry jointly and conditions on the observed surface only softly, so it re-synthesizes and slightly smooths that surface instead of copying it (Sec. 4). Since the observed geometry is already accurate, it can simply be kept. After decoding, we overwrite the predicted TUDF in the visible near-surface band, i.e., the voxels within $b = 1$ voxel of the MoGe2 visible surface, with the visible TUDF computed from the MoGe2 pointmap, the same visible field the model is conditioned on. Both fields live on the same frustum-aligned 256^3 grid by construction, so the merge is a per-voxel replacement with no resampling. Because only this one-voxel band changes, an occluded surface lying more than one voxel behind the visible surface is never affected. The step is training-free and optional; it is not used for the results in the main paper.

Table 8 reports the result on both benchmarks under the evaluation protocol of Tables 1 and 2; the VolFill rows repeat those tables. Reinstatement lowers the visible Chamfer distance to 2.01 on SCRREAM and to 3.56 on NRGB-D. Measured on identical samples, it also slightly improves occluded geometry (occluded CD -0.27 on SCRREAM and -0.15 on NRGB-D), since the more accurate visible points tighten the global Sim(3) alignment used for evaluation, while complete-scene accuracy changes little (complete CD -0.10 on SCRREAM and -0.02 on NRGB-D). We observed no seams between the reinstated band and the generated geometry.

Table 8: **Training-free visible reinstatement** on SCRREAM and NRGB-D ($\times 10^2$).

Method	**Visible**			**Occluded**			**Complete**		
	CD↓	$\text{APD}_{0.1}$ ↑	$\text{APD}_{0.02}$ ↑	CD↓	$\text{APD}_{0.1}$ ↑	$\text{APD}_{0.02}$ ↑	CD↓	$\text{FS}_{0.1}$ ↑	$\text{FS}_{0.02}$ ↑
SCRREAM									
VolFill	2.84	96.10	56.73	3.46	92.70	**55.53**	3.03	95.03	54.83
+ visible reinstatement	**2.01**	**99.70**	**64.99**	**3.30**	**93.62**	54.40	**2.96**	**95.43**	**55.72**
	CD↓	$\text{APD}_{0.1}$ ↑	$\text{APD}_{0.05}$ ↑	CD↓	$\text{APD}_{0.1}$ ↑	$\text{APD}_{0.05}$ ↑	CD↓	$\text{FS}_{0.1}$ ↑	$\text{FS}_{0.05}$ ↑
NRGB-D									
VolFill	4.52	90.06	69.22	6.97	81.41	62.48	6.44	84.35	**63.92**
+ visible reinstatement	**3.56**	**96.05**	**76.66**	**6.34**	**82.90**	**64.85**	**6.40**	**84.67**	63.81

A.5 Additional Quantitative Analysis

Seed stability. Tables 1 and 2 report a single sample per image. Table 9 repeats the SCRREAM evaluation with five sampling seeds, varying only the initial noise (all 460 samples, same protocol). The metrics change little across seeds (coefficient of variation 0.2–2.6%), and the single-seed values of Table 1 lie close to the five-seed means, so the reported results are representative rather than a favorable draw.

Scene extent. Our output grid is a fixed 256^3 box fitted to the camera frustum, so its effective voxel size grows with the scene. To test whether this costs accuracy within the evaluated range, we split the 460 SCRREAM samples into tertiles by the grid's side length and report complete-scene accuracy per bin, under the protocol of Table 1 (Table 10). Although the median voxel size nearly doubles from the small to the large bin (8.7 to 16.8 mm), accuracy does not drop for larger scenes: the large bin scores comparably to the medium one, while the small bin of close-range views is the hardest. Within this range, the 256^3 grid does not limit accuracy; much larger scenes would still call for spatial chunking (Sec. 5).

Table 9: **Seed stability** on SCRREAM ($\times 10^2$): five sampling seeds (mean $\pm$ std) vs. the single seed of Table 1.

	Visible	Occluded		Complete		
	CD↓	CD↓	$APD_{0.1}$ ↑	CD↓	$FS_{0.1}$ ↑	$FS_{0.02}$ ↑
Single seed (Table 1)	2.84	3.46	92.70	3.03	95.03	54.83
Five seeds	2.88 ± 0.06	3.48 ± 0.07	92.77 ± 0.26	3.11 ± 0.08	94.90 ± 0.23	54.20 ± 0.52

Table 10: **Accuracy vs. scene extent** on SCRREAM: complete-scene metrics per tertile of grid side length (CD and FS $\times 10^2$; median voxel size per bin).

Extent	Samples	Grid side (m)	Voxel size (mm)	CD↓	$FS_{0.1}$ ↑	$FS_{0.02}$ ↑
Small	153	1.6–2.5	8.7	4.19	91.70	42.17
Medium	153	2.5–3.4	10.9	2.84	95.77	59.00
Large	154	3.5–5.9	16.8	2.86	96.21	56.34

A.6 Additional Qualitative Results

We refer the reader to our project page, `ngoductuanlhp.github.io/VolFill`*, for interactive 3D visualizations of the reconstructed scenes.*

Comparison with pixel-aligned geometry approaches is presented in Fig. 8. While these baselines [48, 85] achieve high accuracy on visible surfaces, they are inherently restricted to the camera's line of sight, leaving significant holes and "shadows" in occluded regions. In contrast, VolFill synthesizes physically plausible hidden structures, producing a continuous and structurally coherent scene representation resolving the visibility constraints of prior work.

Comparison of mesh reconstruction. As shown in Fig. 9, LaRI and NOVA3R can produce reasonable point clouds in some cases, yet their unstructured outputs consistently yield fragmented, artifact-heavy meshes. Our approach produces cleaner point clouds and topologically consistent meshes that better reflect the physical scene geometry.

Qualitative results of the hybrid 3D VAE are shown in Fig. 10. Our VAE effectively compresses sparse high-resolution TUDF grids into a compact dense latent while preserving fine geometric details upon reconstruction. For larger scenes, a slight reduction in fidelity is observed given the fixed latent capacity, though overall structure and surface topology remain well-preserved—confirming that the latent space provides a faithful scene representation for the downstream diffusion process.

Generative inference trajectory. Fig. 11 illustrates the iterative refinement of the amodal geometry throughout the denoising trajectory. Each panel decodes the model's clean-latent estimate at flow time t, which runs from 0 (noise) to 1 (data) over the 50 Euler steps. Even at an early stage ($t = 0.06$), the model already predicts the coarse overall structure of the scene. As the latent becomes progressively cleaner, the geometry sharpens significantly by $t = 0.30$, followed by continuous, fine-grained refinement until the final step.

Out-of-domain inputs. VolFill is trained only on indoor scenes (3D-FRONT and ScanNet++). Fig. 12 shows its results, without any fine-tuning, on images outside this domain, including failure cases. On ETH3D [73] (Fig. 12a), it still recovers coherent scene layouts with plausible hidden geometry for a semi-outdoor courtyard, a large warehouse (where the reconstruction becomes partly fragmented), and two indoor office spaces. It fails on long-range outdoor street scenes from KITTI [24] and on synthetic, low-light frames from Sintel [7] (Fig. 12b): the geometry fragments and the hidden regions become noisy. These scenes lie far outside the indoor training distribution in scale and appearance; extending VolFill to them would likely require more diverse training data and spatial chunking of the grid (Sec. 5).

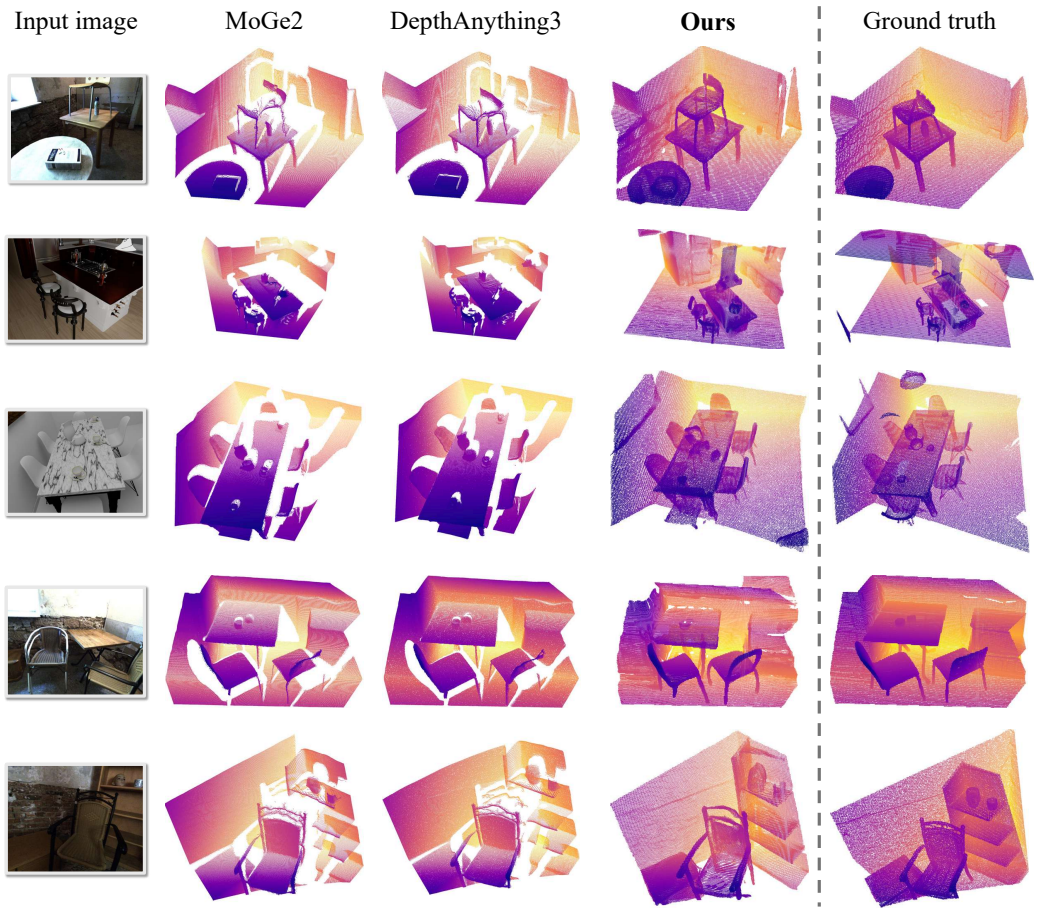


Figure 8: **Qualitative comparison with pixel-aligned approaches.** Unlike MoGe2 [85] and DepthAnything3 [48], which are restricted to visible surfaces and leave significant holes, VolFill reconstructs complete, physically plausible amodal geometry.

B Computing Resources

Our research utilized multiple computing servers, consisting of a high-performance cluster for model training and a dedicated local workstation for data preparation, evaluation, and visualization.

Hardware specifications.

- **Cluster Training:** Primary training was conducted on a cluster node equipped with 2× NVIDIA RTX A6000 GPUs (48 GB VRAM each). We also had access to a SLURM-managed server providing 2× NVIDIA L40S GPUs (48 GB VRAM each).
- **Local Workstation:** Data preprocessing, quantitative evaluation, and visualization were performed on a local workstation featuring an NVIDIA RTX 4090 GPU (24 GB VRAM) and an Intel Core i9 processor with 64 GB of RAM.

Data preparation and training duration. The computational timeline for the final models reported in the main text is as follows:

- **Offline Preprocessing:** The preparation of the training data, including TUDF preparation and the pre-extraction of VAE latents to accelerate training, was performed on the local workstation. This process required approximately 3–4 days for the complete dataset.
- **VAE Training:** The hybrid 3D VAE required approximately 3 days of training using 2 GPUs.

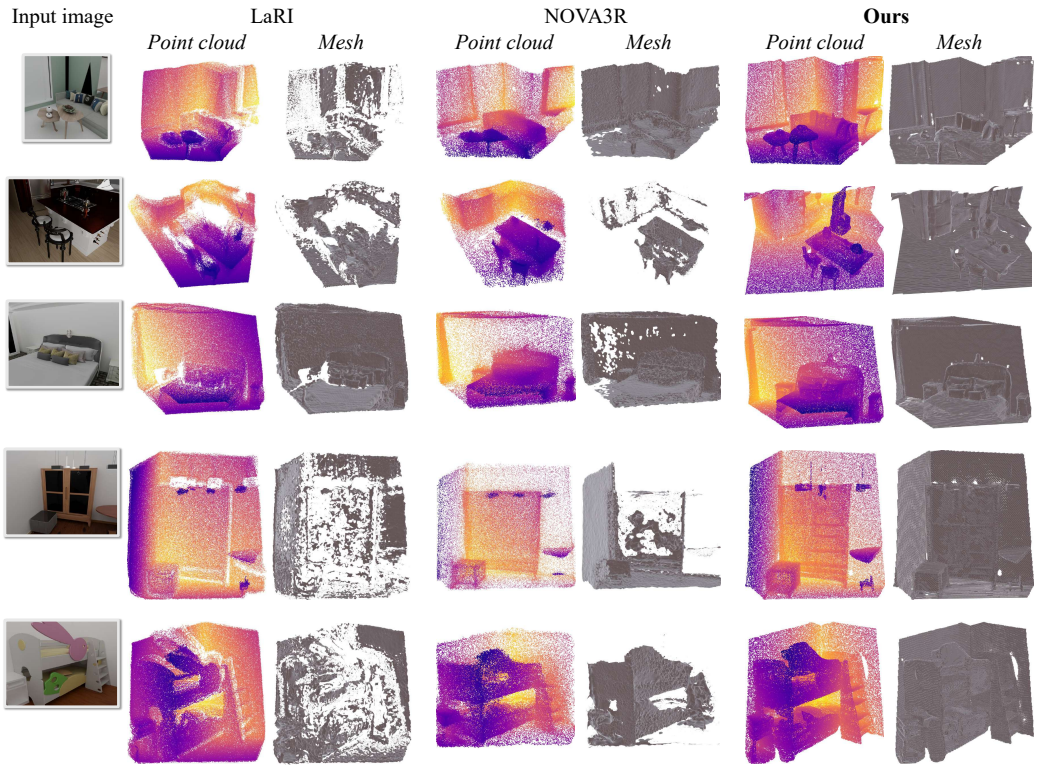


Figure 9: **Point cloud and mesh comparison.** Our method produces cleaner point clouds and significantly more coherent meshes than LaRI and NOVA3R.

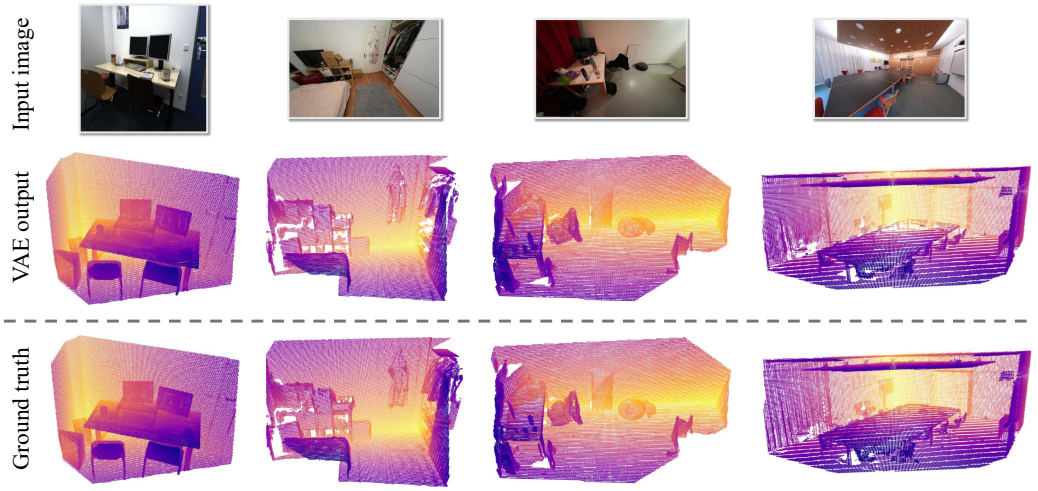


Figure 10: **Qualitative results** of the hybrid 3D VAE.

- **DiT Training:** The latent Diffusion Transformer (DiT) required approximately 2.5 days using 2 GPUs.
- **Ablation Studies:** The dense VAE ablation required approximately 6 days of compute, while DiT conditioning ablations mirrored the 2.5-day schedule.

Total compute expenditure. Accounting for preliminary investigations and hyperparameter tuning, we estimate a total compute expenditure of approximately 1,680 GPU hours (2 GPUs over 35

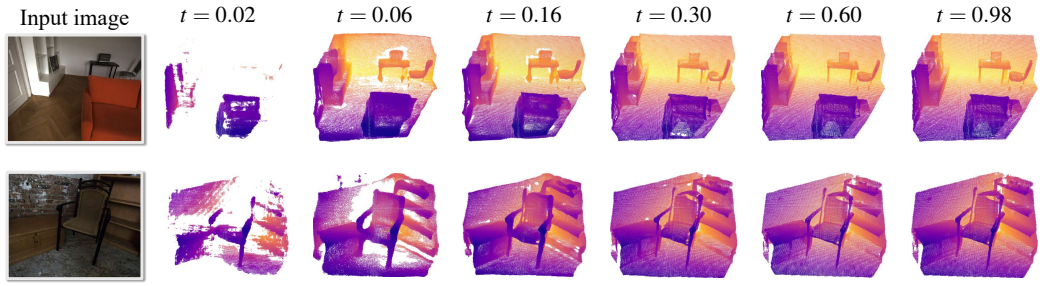
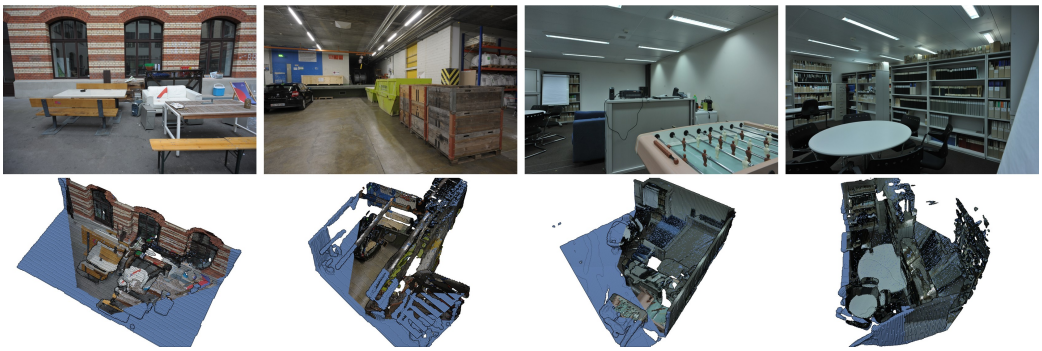
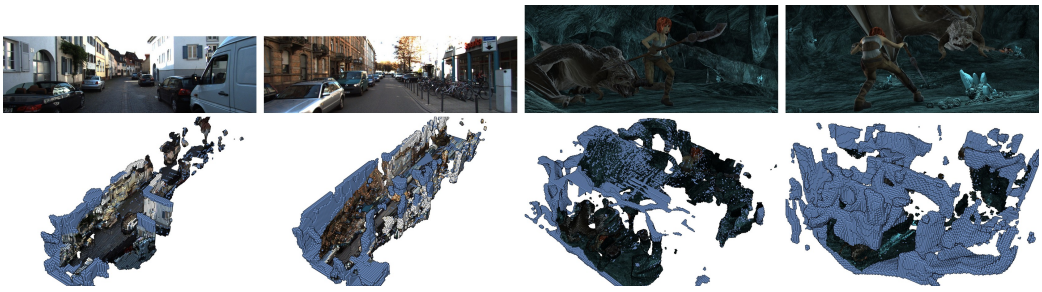


Figure 11: **Evolution of amodal geometry.** Clean-geometry estimates along the denoising trajectory at flow time $t \in (0, 1)$ ($t = 0$: noise, $t = 1$: data). Our model rapidly converges on the coarse scene layout by $t = 0.06$ and recovers sharp, detailed structures by $t = 0.30$, followed by continuous refinement.



(a) Successes: ETH3D courtyard (semi-outdoor), warehouse, and two indoor office spaces.



(b) Failures: KITTI street scenes and Sintel synthetic frames.

Figure 12: **Out-of-domain inputs.** VolFill, trained only on indoor data, applied without fine-tuning to images outside its training domain (input above each reconstruction). As in Fig. 1, geometry seen in the input view keeps the photo's color and hidden geometry recovered by VolFill is shown in **blue**.

days). This excludes the additional CPU-intensive preprocessing and evaluation time on the local workstation.

C Broader Societal Impacts

Our work advances single-image 3D scene reconstruction, which may benefit applications in robotics, assistive navigation, AR/VR, digital twins, architectural design, and embodied AI by enabling richer spatial understanding from limited visual input. In particular, complete-scene reconstruction could help autonomous systems reason about occluded geometry and improve safety in indoor navigation or human-robot interaction. However, the same capability may also raise concerns if deployed in privacy-sensitive settings, since inferred 3D layouts could reveal information about private homes, workplaces, or personal environments beyond what is directly visible in an image. The method could also be misused for surveillance, unauthorized mapping, or synthetic reconstruction of restricted

spaces. In addition, models trained on datasets such as 3D-FRONT and ScanNet++ may inherit dataset biases toward particular indoor layouts, object categories, geographic regions, or socioeconomic settings, potentially limiting performance in underrepresented environments. Thus, we view this work as a research contribution rather than a deployment-ready system, and recommend that practical use include consent-aware data collection and privacy safeguards.