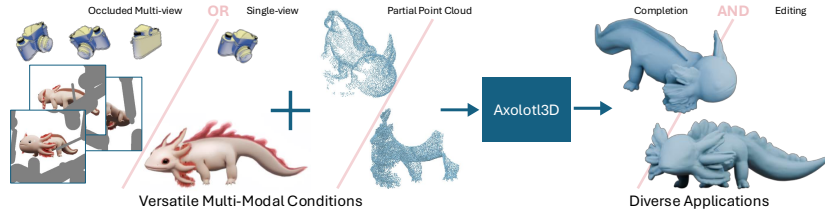


# Axolotl3D: a Unified Framework for Faithful 3D Shape Completion

Anita Hu<sup>1</sup>  and Maria Shugrina<sup>1</sup> 

NVIDIA

<https://research.nvidia.com/labs/sil/projects/axolotl3d>



**Fig. 1:** Axolotl3D enables controllable, occlusion-aware 3D generation across diverse applications with consistent and faithful shape completion.

**Abstract.** Recent 3D generative models produce high-quality geometry from a single image using large-scale priors and diffusion architectures. However, they assume complete visibility and single-view inputs, limiting applicability in multi-view, occluded, or editing scenarios. Although prior works address these challenges individually, they lack a unified framework for controllable 3D completion under diverse conditioning signals.

We present Axolotl3D, a multi-modal and occlusion-aware 3D generation model that jointly conditions on images, visibility masks, camera parameters, and a partial point cloud. The point cloud serves as a geometric anchor promoting faithful shape completion, while camera parameters ensure consistent multi-view alignment in a shared 3D coordinate system. A unified training strategy synthesizes diverse conditioning regimes from large-scale 3D data, enabling robust cross-modal reasoning.

Experiments on Toys4K and OmniObject3D demonstrate state-of-the-art performance under both clean and occluded settings, as well as strong results in real-world reconstruction and geometry-consistent editing.

**Keywords:** Shape Completion · Multi-Modal Generation · 3D Editing

## 1 Introduction

Recent advances in 3D generative modeling, leveraging large-scale 3D datasets and diffusion-based architectures (e.g., Hunyuan3D 2.1 [20], Trellis [43,44]), have enabled detailed and visually compelling geometry synthesis from a single image. These approaches demonstrate the power of learned 3D priors trained on large collections of synthetic and scanned objects. However, despite their strong generative capabilities, such methods remain limited in real-world scenarios, where

objects are often captured under sparse multi-view setups, partially occluded, or require geometry-aware editing of existing meshes. By assuming complete visibility and single-view inputs, existing approaches are unable to reason under occlusion, integrate multiple views, or enforce geometric constraints, restricting their applicability to realistic capture and content creation pipelines.

Several research directions address these challenges in isolation. Works such as Amodal3R [41] focus on recovering occluded geometry, while ReconViaGen [4] emphasizes cross-view consistency. Editing methods including Instant3dit [1], PrEditor3D [10] enable precise user-guided shape modification. More recent approaches, such as AmodalGen3D [53], begin to combine multi-view reasoning with occlusion handling. While effective within their respective scopes, these approaches remain fragmented, limiting their applicability across broad conditioning and editing scenarios. We argue instead that these problems share a common underlying structure: each requires completing a 3D object from incomplete observations while enforcing consistency with explicit geometric constraints. From this perspective, they can be viewed as different manifestations of constrained 3D completion, motivating a unified and controllable generation framework.

We introduce Axolotl3D, a multi-modal and occlusion-aware 3D generation model that conditions jointly on images, per-view visibility masks, camera parameters, and points. To promote geometric faithfulness, a partial point cloud—sampled from mesh surfaces or projected from predicted depths—acts as a geometric anchor. Camera parameters help align these 3D points and multi-view images within a shared coordinate system, allowing the model to integrate visual and geometric evidence consistently. To enable robust cross-modal reasoning, we further propose a unified training strategy that synthesizes diverse conditioning regimes from large-scale 3D mesh datasets. By simulating partial observations, occlusions, and mixed image–geometry inputs, we train the model to generalize across single-view, sparse multi-view, and editing scenarios.

We evaluate our approach on 3D object datasets including Toys4K [36] and OmniObject3D [42] under both single-view and multi-view settings, with and without synthetic occlusions. Across all scenarios, Axolotl3D outperforms state-of-the-art methods in geometric accuracy and reconstruction fidelity. Furthermore, we demonstrate its applicability in real-world pipelines by reconstructing objects from images using predicted points and cameras from MapAnything [22] and performing geometry-consistent object editing using views inpainted by Stable Diffusion Inpainting [33].

In summary, our contributions are threefold:

- We introduce a unified multi-modal, occlusion-aware 3D generation framework that supports single-view generation, sparse multi-view reconstruction, occlusion completion, and geometry-aware editing within a single model.
- We propose a cross-modal training strategy that simulates diverse conditioning scenarios from large-scale 3D data, enabling robust reasoning under partial and heterogeneous observations.

- We achieve state-of-the-art performance on benchmark datasets, producing faithful geometry and effectively handling occlusions, while demonstrating effectiveness in real-world applications.

## 2 Related Work

**Multi-View 3D Generation:** While recent image-to-3D generative models [20, 43, 44] achieve impressive mesh quality from a single image, fewer works address sparse, unposed multi-view shape generation. Although single-view models can be adapted to multi-view inputs—for example, by cycling views during denoising as in Trellis [44]—without jointly modeling cross-view interactions, geometric consistency remains limited. An alternative two-stage strategy first performs novel view synthesis (NVS) before 3D generation: EscherNet [25] produces consistent novel views from one or more reference images for subsequent reconstruction with NeuS [39], while SpaRP [45] outputs six fixed views for 3D generation with One-2-3-45++ [27]. These pipelines can be brittle, as small inconsistencies in the synthesized views may compromise fidelity and detail. With the emergence of large-scale 3D foundation models [22, 38, 40], we can now directly predict explicit 3D attributes from images—such as camera parameters, depth maps, and point maps—providing strong priors for downstream tasks. VGGT [38] demonstrates impressive cross-view consistency in novel view synthesis, with its features leveraged by ReconViaGen [4] for accurate multi-view generation. Motivated by this progress, we condition our generative model on explicit geometric signals—point clouds and camera parameters—allowing flexible application to posed and unposed multi-view scenarios. A concurrent work ShapeR [34] also generates 3D shapes from multi-view images, sparse point clouds, and camera poses, but is tailored to SLAM [9] outputs.

**Geometry-Controlled Generation:** Some recent 3D generation methods incorporate geometric cues to guide the generated shape. 3DShape2VecSet [50] introduces a fixed-length latent representation (VecSet) for shapes and demonstrates applications including 3D generation from partial point clouds. CLAY [51] supports diverse conditioning modalities, including sparse points, voxels, bounding boxes, and multi-view images, but it is not designed to handle partial point clouds or occluded views. PoseMaster [46] generates articulated 3D characters conditioned on skeletons and images. More recently, Hunyuan3D-Omni [21] unifies geometry-controlled generation across point clouds, bounding boxes, voxels, and skeleton inputs, though it only accepts a single image and cannot robustly handle incomplete observations. In contrast, our method handles both multi-view images and point cloud inputs, supporting occlusions in both modalities.

**Amodal Object Completion:** Amodal object completion aims to recover the full 3D shape from partial observations, typically images and segmentation masks. A related line of work focuses on single-image generation, producing complete 3D scenes and objects from monocular inputs. Methods such as CAST [47]

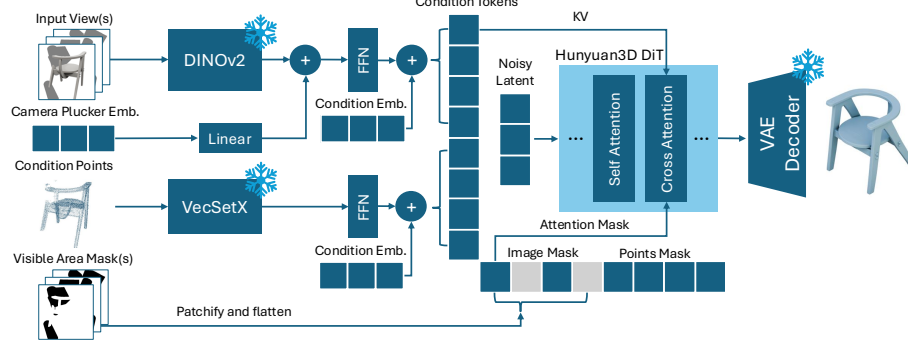
and SAM 3D [5] generate object-level reconstructions and align them within the scene, using monocular depth estimators for partial point clouds and orientation cues. MIDI [19] generates all objects in a scene jointly, allowing the model to capture inter-object relationships and maintain spatial coherence without requiring a separate layout optimization step. At the object level, Amodal3R [41] reconstructs occluded geometry from a single image by leveraging occlusion and visibility masks, using mask-weighted multi-head cross-attention and occlusion-aware conditioning layers. Occlusion-aware view synthesis methods, such as EscherNet++ [52] and DeOcc-1-to-3 [31], generate novel views from occluded inputs to facilitate amodal completion. However, the reconstructed geometry often lacks fine details, revealing limitations of NVS-based pipelines. A recent work, AmodalGen3D [53], extends Amodal3R [41] to unposed multi-view input and incorporating partial points from  $\pi^3$  [40], producing more geometrically faithful reconstructions. In comparison, our method targets the posed setting, yet show strong results in an unposed ablation.

**Editing:** Methods perform 3D editing using different strategies, from multi-view inpainting to latent-space mesh manipulation. Instant3dit [1] and PrEditor3D [10] perform text-guided edits through multi-view image inpainting from four fixed viewpoints, followed by 3D reconstruction. ObjFiller-3D [11] extends this paradigm using video diffusion, removing the restriction on viewpoints and improving cross-view consistency. More recently, concurrent work VecSet-Edit [18] enables localized mesh editing guided by a single image directly in the VecSet latent space, better preserving appearance in untouched regions. While effective for targeted edits, these methods cannot handle general object completion without an initial object or guidance signal.

### 3 Method

We consider the task of 3D object completion from partial and heterogeneous observations, common in scenarios like 3D capture or editing. For multi-view reconstruction, the known information typically contains multiple occluded views and partial geometry, e.g. after processing an image sequence with techniques like [13, 23, 38]. For targeted shape editing, completion should respect fixed geometry regions and take additional guidance, e.g. an edited view of the object. To support such applications, our goal is to predict a complete and geometrically consistent 3D shape given: a collection of images  $\{I_i\}_{i=1}^N$ , corresponding masks  $\{M_i\}_{i=1}^N$  and camera parameters  $\{C_i\}_{i=1}^N$ , and a partial point cloud  $P$ . Each mask  $M_i$  specifies regions of  $I_i$  that provide valid observations. For amodal completion,  $M_i$  encodes visibility, whereas for object editing, it indicates regions to preserve while users provide modified views with fully valid  $M_i$  to drive geometry updates. By treating all visual, geometric, and user-guided signals as partial observations of the object, our formulation unifies single- and multi-view generation, amodal completion, and geometry-aware editing under a single framework.





**Fig. 2: Overview of Axolotl3D.** Given multi-modal inputs—posed images, visibility masks, and partial points—our method encodes each modality and fuses them into multi-modal condition tokens. We fine-tune Hunyuan3D-DiT on these tokens to produce completed shape latents, which are decoded by the ShapeVAE.

### 3.1 Model Architecture

Axolotl3D is built upon recent state-of-the-art image-to-3D diffusion model, Hunyuan3D 2.1 [20]. Its shape generation model comprises a shape autoencoder (Hunyuan3D-ShapeVAE) that encodes meshes into a compact latent representation, and an image-conditioned flow-based diffusion transformer (Hunyuan3D-DiT) trained to predict shape tokens in this latent space, which are then decoded into meshes. We adapt this DiT backbone to jointly process all heterogeneous inputs, replacing the original single-image cross-attention with a multi-modal cross-attention mechanism that integrates visual and geometric cues simultaneously. An overview of the proposed architecture is shown in Fig. 2.

**Visual Encoding:** Following original architecture [20], each input image  $I_i$  is encoded into  $\text{enc}(I_i)$  using DINOv2 [29], resulting in rich spatially-aligned visual features suitable for multi-view reasoning. Our model takes up to  $N = 6$  images, where missing views are zero-padded and masked. To enable cross-view alignment of visual features in a shared 3D coordinate system, our model also ingests corresponding camera parameters  $C_i$ , represented as  $\text{pl}(C_i)$  using Plücker embeddings [35]. Specifically, for each pixel  $(u, v)$ , its Plücker embedding is:

$$\mathbf{p}_{u,v} = (\mathbf{o} \times \mathbf{d}_{u,v}, \mathbf{d}_{u,v}) \in \mathbb{R}^6, \quad (1)$$

where  $\mathbf{o}$  is the camera center and  $\mathbf{d}_{u,v}$  is the corresponding ray direction from camera center to the pixel. These  $\text{pl}(C_i)$  are first average-pooled to match the spatial resolution of the image features  $\text{enc}(I_i)$ , and then projected via a linear layer to match their channel dimension, resulting in  $\text{enc}(C_i)$ . We obtain final visual embedding  $\text{enc}_V(I_i, C_i)$  by adding  $\text{enc}(C_i)$  to  $\text{enc}(I_i)$ , enriching image features with camera information. Each mask  $M_i$  is partitioned into non-overlapping  $14 \times 14$  patches (patch size of DINOv2), where each patch corresponds to a pixel in the downscaled mask  $B_i$  matching the resolution of  $\text{enc}(I_i)$ . A pixel in  $B_i$  is set to 1 if it is completely unoccluded and 0 otherwise. Note that the background

is treated as unoccluded to enforce free-space constraints. These masks  $B_i$  bias the attention toward unoccluded regions (see later), similar to Amodal3R [41].

**Geometry Encoding:** To provide additional geometric grounding and ensure faithful completions, partial point cloud  $P$  forms an additional conditioning input. We encode  $P$  into  $\text{enc}(P)$  using VecSetX [49], a follow-up to [50], which maps input points into a compact, fixed-length set of latent codes. Pretrained on ShapeNet-v2 [3], it yields high fidelity reconstruction retaining local details.

**Multi-Modal Fusion:** We now process geometric features  $\text{enc}(P)$  and camera-enriched multi-view features  $\{\text{enc}_V(I_i, C_i)\}_{i=1}^N$  into multi-modal feature tokens  $F$ . Each modality’s features are flattened into point tokens  $F_p \in \mathbb{R}^{T_p \times C_p}$  and combined visual tokens  $F_v \in \mathbb{R}^{NT_v \times C_v}$ , where  $N$  is number of views (6),  $T_v$  and  $T_p$  are the token lengths and  $C_v$  and  $C_p$  are the channel dimensions of each modality. Then, each token in  $F_p$  and  $F_v$  is processed with its respective per-modality gated feedforward network (FFN) of similar design used in [5, 14] (see Supplement), mapping tokens to a unified channel dimension  $C = 1024$ . This produces richer, modality-specific features that emphasize important signals and prepare tokens for multi-modal fusion. Two learnable embeddings, one per modality (visual and geometric) of shape  $C$  are added element-wise to each token of the corresponding modality to differentiate between visual and geometric inputs, resulting in final tokens  $F'_p \in \mathbb{R}^{T_p \times C}$  and  $F'_v \in \mathbb{R}^{NT_v \times C}$ .

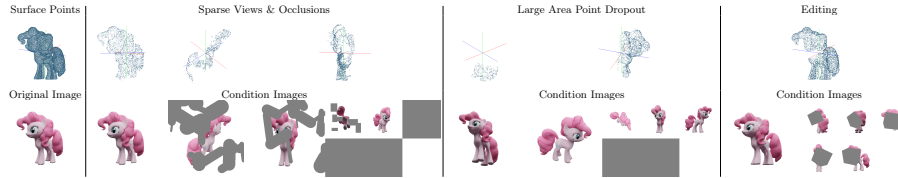
**Mask-biased Cross-Attention:** We concatenate  $F'_p$  and  $F'_v$  along the token dimension, forming a set of multi-modal condition tokens  $F$  of length  $T = T_p + NT_v$  and channel dimension  $C$ , which are used in the multi-head cross-attention modules of our DiT, trained to denoise the shape latent  $Z \in \mathbb{R}^{L \times C'}$  across timesteps (See [16] for background). Let  $Q = ZW_Q$ ,  $K = FW_K$ ,  $V = FW_V$  be the query, key, and value projections of the latent and condition tokens. For multi-head cross-attention with  $H$  heads, the attention is computed as:

$$\text{CrossAttn} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{D}} + B\right)V, \quad (2)$$

where each head has dimension  $D = C'/H$ , and  $B \in \mathbb{R}^{L \times T}$  is an attention bias derived from the patchified image masks  $B_i$ . Specifically, entries corresponding to occluded visual tokens are set to  $-\infty$ , preventing them from contributing to attention, while unoccluded tokens are assigned 0 (unbiased). Point tokens, which are unmasked due to their fixed-length representation, are also set to 0.

### 3.2 Synthetic Data Augmentation

We train our model on a large collection of textured 3D meshes, synthesizing diverse conditioning regimes for robust cross-modal reasoning. During dataset preprocessing, we normalize each mesh  $\mathcal{M}_k$  to a unit cube  $[-1, 1]^3$  and sample



**Fig. 3: Data Augmentation:** Examples of our on-the-fly data augmentation.

a dense surface point cloud  $P_{\text{surface}}^k$  (see Supplement). We also sample 150 cameras  $C^k$  on a sphere and render corresponding images  $I^k$  following [44]. During training, we sample and augment this data to produce input partial point cloud  $P$  and multi-view  $\{I_i\}_{i=1}^N$ ,  $\{M_i\}_{i=1}^N$ ,  $\{C_i\}_{i=1}^N$ ,  $N = 6$ , according to the following scenarios chosen with equal probability (if desired, this can be adjusted to prioritize specific regimes). See Fig. 3 for an example of each augmentation.

**Sparse Views & Occlusions:** To support amodal completion and promote learning of cross-modal correspondences, we simulate sparse view observation settings with varying levels of occlusion. We select one random view  $I_0 \in I^k$  as the unobserved view, and sample up to  $N$  conditioning cameras  $C_i \in C^k$  from non-nearby views defined based on the angle between camera backward vectors (with  $90^\circ$  angle threshold). Each sampled view  $I_i$  is occluded with probability  $p_{\text{occl}}$ . When occlusion occurs, we generate a random mask  $\tilde{M}_i$  using brush strokes or bounding boxes following LaMa [37]. We then remove points in  $P_{\text{surface}}^k$  that are not visible in the sampled views or project into regions covered by occlusion masks (see Supplement for details), resulting in realistic patchy point clouds. In addition, we further introduce a masking probability  $p_{\text{mask}}$ , to avoid over-reliance on point-based cues and promote complementary multi-modal learning. With this probability, the same occlusion mask  $\tilde{M}_i$  is applied to the multi-view inputs  $I_i$  and their masks  $M_i$ . Using this procedure, we can further divide this task into single- and multi-view settings, sampled equally during training. We set  $p_{\text{mask}} = 0.5$  for both settings. In the single-view setting, where many points are missing due to limited visibility, we use  $p_{\text{occl}} = 0.25$ , while in the multi-view setting we set  $p_{\text{occl}} = 1.0$ .

**Large Area Point Dropout:** This augmentation simulates scenarios where large regions of the object are missing and unobserved from camera viewpoints, requiring the model to infer missing geometry from limited points and indirect image observations. This provides more extreme point dropout compared to the previous, training the model to not over-rely on available points. First, we randomly select a dropout axis ( $x$ ,  $y$ , or  $z$ ) and a direction (positive or negative), then sample a threshold along this axis. Points in  $P_{\text{surface}}^k$  beyond the threshold in the chosen direction are removed, with the drop probability decreasing linearly with distance from the threshold, creating a smooth gradient of missing geometry and more natural partial observations. Afterwards, points are further filtered based on the alignment of their normals with the dropout direction mimicking realistic visibility effects, with the drop probability proportional to the

cosine similarity between the normal and the downward direction. To prevent the missing region from being directly observed, which is more realistic in real-world sparse-view scenarios, we filter out camera views whose backward vectors are within  $90^\circ$  of the dropout direction. The final set of conditioning views are then sampled from the remaining cameras. Points not visible in any conditioning camera view are removed using their projected locations and depth comparisons.

**Editing:** In local editing scenarios, a bounding box or other simple geometric primitives are often used to specify the region to be modified, as in the 3D interface of Instant3dit [1], and edits are typically guided by a reference image or a text prompt [1, 10, 18]. In this work, we focus on image-based guidance, where additional views may be provided to help preserve fine details in the unedited regions and maintain cross-view consistency. We emulate such user-guided shape modification by selecting a random camera view as the "modified" view. We then place a 3D bounding box centered on a randomly chosen point in  $P_{\text{surface}}^k$  that is visible in this view. All points inside the bounding box are removed from  $P_{\text{surface}}^k$ , creating a region for inpainting. Additional camera views are then randomly sampled to provide observations of the unmodified regions, with the edited region masked in all  $I_i$ ,  $M_i$  by projecting the bounding box using camera  $C_i$  (see Supplement for details). This creates a task where only the modified view shows the full object with the edit, while the remaining views reveal only the surrounding geometry outside the edited region.

## 4 Experiments

### 4.1 Data and Setting

**Training Data:** We train on 407k shapes from TRELLIS-500K [44], containing aesthetics-filtered objects from ObjaverseXL [8], ABO [6], 3D-FUTURE [12], and HSSD [24]. We remove buildings and scenes from the ObjaverseXL subset using a classifier trained on G-Objaverse labels [54], as these often depict multi-object scenes (see Supplement). Shapes are scaled and rendered with turn-around cameras at fixed distance and intrinsics [44]. The geometry is preprocessed following Hunyuan3D 2.1 [20], including normalization, watertight processing, SDF and surface sampling. Train-time augmentations are described in §3.2.

**Implementation:** For our backbone we adopt Hunyuan3D 2.1 [20] and fine-tune from pre-trained weights utilizing their strong 3D generation priors. Additional layers and embeddings in §3.1 are trained from scratch: linear layers use Kaiming uniform initialization [15], while embeddings use normal initialization  $\mathcal{N}(0, 1/\sqrt{n})$  where  $n$  is the condition embedding dimension (1024). We use classifier-free guidance (CFG) [17] with drop rate of 0.1 and AdamW [28] optimizer with learning rate of  $1e-5$ . The model is trained for 500k steps with batch size of 16 on 8 A100 GPUs. At inference, the CFG guidance scale is set to 5 and sampling steps is 50.

**Evaluation Data:** For quantitative evaluation (§4.3), we use Toys4K [36] and OmniObject3D [42]. Toys4K contains 4k synthetic objects in 105 categories, and OmniObject3D has 6k scanned objects in 190 categories. To simulate realistic 3D-consistent occlusions, we add random occluders from a set of shape primitives (see Supplement). Different from training, we obtain point condition by backprojection from rendered depth maps of each sampled view, excluding occluded regions. This is downsampled to 8,192 points for ShapeR and our method while Hy3D-Omni supports a max of 2,048 points and SAM3D uses pointmaps of shape 518x518x3. We further provide comparisons using 2,048 input points and evaluate robustness using predicted depth from Depth Anything v3 [26], both with and without camera perturbations. Predicted depth is aligned to ground-truth depth via per-scene scale and shift, then backprojected using ground-truth camera parameters. Camera extrinsics are perturbed with Gaussian noise ( $10^\circ$  rotation, 10% translation scale), while the focal length is jittered by  $\pm 5\%$ . For qualitative visualization (§4.4), we focus on Toys4K’s geometrically challenging objects with richer appearance details (OmniObject3D results in Supplement). For applications (§4.6), we show single-view generation with MapAnything [22] using MVImgNet images [48], and editing results on Toys4K objects.

## 4.2 Baselines and Metrics

**Baselines:** We compare our method to the recent state-of-the-art methods [41], [21], [5], and to one concurrent work [34]. We omit AmodalGen3D [53] as its code is not publicly available. Amodal3R [41] is capable of single- and multi-view completion and is the only method that does not accept point condition. Hunyuan3D-Omni (Hy3D-Omni) [21] and SAM 3D [5] are both single-view methods that accept point condition, where SAM 3D pointmaps are in camera space. Concurrent ShapeR [34] is the only multi-view method that also accepts points. Because Hy3D-Omni is not trained for occlusions, we complete partial views with Pix2Gestalt [30] before inference. While it accepts only a single image, points from multiple views can still be used. Thus, we evaluate all models, except SAM 3D, on multi-view and single-view conditioned generation. In the multi-view scenario, Hy3D-Omni is given the best view out of the available views as image input (selected based on max object coverage; see Supplement). Finally, to provide required prompts to ShapeR, we use captions included with TRELLIS-500K [44] for Toys4k [36] and object category for OmniObject3D [42].

**Metrics:** To assess the fidelity and completeness of the reconstructed shapes, we choose complementary geometric metrics. F-score (at 0.05 threshold) measures point-level correspondence, balancing precision and recall to assess accuracy and coverage. Voxel IoU (vIoU) at  $64^3$  resolution evaluates coarse volumetric agreement, sensitive to major occupancy errors. Chamfer Distance (CD) captures fine-grained geometric deviations and penalizes missing or distorted regions. Because not all methods use world-space points as input, we first align generated and groundtruth meshes, normalized to  $[-1, 1]$ , using ICP [2]. 1M points are uniformly sampled from both meshes before computing the metrics.

**Table 1: Quantitative Evaluation**, marking **best** and **second-best**. Our method shows superior performance against state-of-the-art and performs comparably to ShapeR, with lower variance and stronger single-view results.

| Scenario:<br>Method         | Toys4k                 |                        |                             |                        |                        |                             | OmniObject3D           |                        |                             |                        |                        |                             |
|-----------------------------|------------------------|------------------------|-----------------------------|------------------------|------------------------|-----------------------------|------------------------|------------------------|-----------------------------|------------------------|------------------------|-----------------------------|
|                             | Without Occlusion      |                        |                             | With Occlusion         |                        |                             | Without Occlusion      |                        |                             | With Occlusion         |                        |                             |
|                             | F-score $\uparrow$     | vIoU $\uparrow$        | CD $\downarrow_{\times 10}$ | F-score $\uparrow$     | vIoU $\uparrow$        | CD $\downarrow_{\times 10}$ | F-score $\uparrow$     | vIoU $\uparrow$        | CD $\downarrow_{\times 10}$ | F-score $\uparrow$     | vIoU $\uparrow$        | CD $\downarrow_{\times 10}$ |
| <b>Single-View:</b>         |                        |                        |                             |                        |                        |                             |                        |                        |                             |                        |                        |                             |
| Amodal3R [41]               | 0.7310<br>$\pm 0.2365$ | 0.2129<br>$\pm 0.1139$ | 0.5460<br>$\pm 0.6305$      | 0.6877<br>$\pm 0.2381$ | 0.1886<br>$\pm 0.1063$ | 0.6144<br>$\pm 0.6495$      | 0.4847<br>$\pm 0.2646$ | 0.1459<br>$\pm 0.0897$ | 1.1406<br>$\pm 0.9197$      | 0.4724<br>$\pm 0.2587$ | 0.1382<br>$\pm 0.0845$ | 1.1174<br>$\pm 0.9336$      |
| SAM 3D [5]                  | 0.7160<br>$\pm 0.2491$ | 0.2039<br>$\pm 0.1049$ | 0.5672<br>$\pm 0.6464$      | 0.6764<br>$\pm 0.2504$ | 0.1836<br>$\pm 0.1018$ | 0.6375<br>$\pm 0.6871$      | 0.4808<br>$\pm 0.2715$ | 0.1440<br>$\pm 0.0932$ | 1.1984<br>$\pm 0.9989$      | 0.4666<br>$\pm 0.2709$ | 0.1407<br>$\pm 0.0932$ | 1.2409<br>$\pm 1.0098$      |
| Hy3D-Omni [21]              | 0.7811<br>$\pm 0.2065$ | 0.2496<br>$\pm 0.1471$ | 0.4491<br>$\pm 0.4261$      | 0.6593<br>$\pm 0.2614$ | 0.1802<br>$\pm 0.1297$ | 0.8123<br>$\pm 0.8209$      | 0.7401<br>$\pm 0.1984$ | 0.2007<br>$\pm 0.1147$ | 0.4913<br>$\pm 0.4132$      | 0.6028<br>$\pm 0.2386$ | 0.1406<br>$\pm 0.1036$ | 0.9018<br>$\pm 0.7425$      |
| ShapeR [34]<br>(concurrent) | 0.7920<br>$\pm 0.2127$ | 0.2573<br>$\pm 0.1235$ | 0.3981<br>$\pm 0.4121$      | 0.7453<br>$\pm 0.2249$ | 0.2285<br>$\pm 0.1208$ | 0.4622<br>$\pm 0.4321$      | 0.7135<br>$\pm 0.2210$ | 0.2160<br>$\pm 0.1008$ | 0.4820<br>$\pm 0.3853$      | 0.6816<br>$\pm 0.2306$ | 0.2012<br>$\pm 0.0961$ | 0.5338<br>$\pm 0.4236$      |
| <b>Ours</b>                 | 0.9221<br>$\pm 0.1044$ | 0.3470<br>$\pm 0.0988$ | 0.2157<br>$\pm 0.1820$      | 0.9046<br>$\pm 0.1140$ | 0.3265<br>$\pm 0.1037$ | 0.2402<br>$\pm 0.2079$      | 0.8716<br>$\pm 0.1187$ | 0.2792<br>$\pm 0.0856$ | 0.2854<br>$\pm 0.2269$      | 0.8613<br>$\pm 0.1191$ | 0.2724<br>$\pm 0.0848$ | 0.2971<br>$\pm 0.2228$      |
| <b>Multi-View (6):</b>      |                        |                        |                             |                        |                        |                             |                        |                        |                             |                        |                        |                             |
| Amodal3R [41]               | 0.8134<br>$\pm 0.2012$ | 0.2516<br>$\pm 0.1236$ | 0.3935<br>$\pm 0.5012$      | 0.7529<br>$\pm 0.2198$ | 0.2166<br>$\pm 0.1154$ | 0.4999<br>$\pm 0.6021$      | 0.5250<br>$\pm 0.2741$ | 0.1585<br>$\pm 0.0968$ | 1.0733<br>$\pm 0.9534$      | 0.5046<br>$\pm 0.2667$ | 0.1484<br>$\pm 0.0905$ | 1.1127<br>$\pm 0.9520$      |
| Hy3D-Omni [21]              | 0.9211<br>$\pm 0.1138$ | 0.3860<br>$\pm 0.1587$ | 0.2111<br>$\pm 0.2108$      | 0.8920<br>$\pm 0.1482$ | 0.3355<br>$\pm 0.1581$ | 0.2738<br>$\pm 0.3667$      | 0.9549<br>$\pm 0.0797$ | 0.3638<br>$\pm 0.1397$ | 0.1677<br>$\pm 0.1580$      | 0.9353<br>$\pm 0.1043$ | 0.3314<br>$\pm 0.1377$ | 0.2005<br>$\pm 0.2317$      |
| ShapeR [34]<br>(concurrent) | 0.9433<br>$\pm 0.1303$ | 0.4476<br>$\pm 0.1449$ | 0.1649<br>$\pm 0.3096$      | 0.9190<br>$\pm 0.1528$ | 0.4002<br>$\pm 0.1553$ | 0.2040<br>$\pm 0.3416$      | 0.9629<br>$\pm 0.0985$ | 0.4412<br>$\pm 0.1318$ | 0.1329<br>$\pm 0.2112$      | 0.9532<br>$\pm 0.0989$ | 0.4123<br>$\pm 0.1313$ | 0.1443<br>$\pm 0.1842$      |
| <b>Ours</b>                 | 0.9768<br>$\pm 0.0445$ | 0.4424<br>$\pm 0.1077$ | 0.1351<br>$\pm 0.0543$      | 0.9689<br>$\pm 0.0534$ | 0.4187<br>$\pm 0.1121$ | 0.1448<br>$\pm 0.0685$      | 0.9842<br>$\pm 0.0425$ | 0.4127<br>$\pm 0.0903$ | 0.1389<br>$\pm 0.1240$      | 0.9755<br>$\pm 0.0473$ | 0.3967<br>$\pm 0.0889$ | 0.1467<br>$\pm 0.1030$      |

**Table 2: Quantitative results with single-view predicted depth** on Toys4k.

| Method                 | w/o Camera Perturb     |                        |                             | w/ Camera Perturb      |                        |                             |
|------------------------|------------------------|------------------------|-----------------------------|------------------------|------------------------|-----------------------------|
|                        | F-score $\uparrow$     | vIoU $\uparrow$        | CD $\downarrow_{\times 10}$ | F-score $\uparrow$     | vIoU $\uparrow$        | CD $\downarrow_{\times 10}$ |
| SAM3D                  | 0.7062<br>$\pm 0.2538$ | 0.1993<br>$\pm 0.1049$ | 0.5898<br>$\pm 0.6643$      | N/A                    | N/A                    | N/A                         |
| Hy3D-Omni              | 0.6837<br>$\pm 0.2170$ | 0.1729<br>$\pm 0.1249$ | 0.5926<br>$\pm 0.4585$      | N/A                    | N/A                    | N/A                         |
| ShapeR<br>(concurrent) | 0.5626<br>$\pm 0.2408$ | 0.1407<br>$\pm 0.0847$ | 0.7767<br>$\pm 0.5569$      | 0.5605<br>$\pm 0.2398$ | 0.1409<br>$\pm 0.0863$ | 0.7750<br>$\pm 0.5518$      |
| <b>Ours</b>            | 0.7616<br>$\pm 0.1731$ | 0.1903<br>$\pm 0.0908$ | 0.4235<br>$\pm 0.3135$      | 0.7620<br>$\pm 0.1712$ | 0.1901<br>$\pm 0.0908$ | 0.4230<br>$\pm 0.3138$      |

**Table 3: Quantitative results with 2048 input points** on Toys4k with occlusions.

| Method                 | Single-View            |                        |                             | Multi-View             |                        |                             |
|------------------------|------------------------|------------------------|-----------------------------|------------------------|------------------------|-----------------------------|
|                        | F-score $\uparrow$     | vIoU $\uparrow$        | CD $\downarrow_{\times 10}$ | F-score $\uparrow$     | vIoU $\uparrow$        | CD $\downarrow_{\times 10}$ |
| Hy3D-Omni              | 0.6594<br>$\pm 0.2614$ | 0.1802<br>$\pm 0.1297$ | 0.8119<br>$\pm 0.8208$      | 0.8920<br>$\pm 0.1482$ | 0.3356<br>$\pm 0.1581$ | 0.2738<br>$\pm 0.3667$      |
| ShapeR<br>(concurrent) | 0.6956<br>$\pm 0.2317$ | 0.2004<br>$\pm 0.1106$ | 0.5377<br>$\pm 0.4505$      | 0.9126<br>$\pm 0.1466$ | 0.3734<br>$\pm 0.1505$ | 0.2119<br>$\pm 0.3078$      |
| <b>Ours</b>            | 0.8738<br>$\pm 0.1291$ | 0.2766<br>$\pm 0.0958$ | 0.2838<br>$\pm 0.2303$      | 0.9653<br>$\pm 0.0568$ | 0.3996<br>$\pm 0.1103$ | 0.1521<br>$\pm 0.0696$      |

### 4.3 Quantitative Results

We present quantitative results in Tb. 1. Under single-view conditions, our method significantly outperforms the baselines including concurrent work, ShapeR. In multi-view conditions, our method shows superior performance against state-of-the-art and comparable with ShapeR. In all scenarios, our method achieves lower standard deviation showing more consistent and stable performance.

We evaluate robustness to realistic inputs using predicted monocular depth without occlusions in Tb. 2. Our method achieves the best geometric fidelity across all metrics, outperforming the baselines. Compared to the ground-truth depth setting in Tb. 1, the performance gap narrows under predicted depth, with our method, ShapeR, and Hy3D-Omni showing larger performance drops, reflecting the sensitivity of geometry-guided reconstruction to errors in monocular depth estimation. In contrast, SAM3D being trained directly on predicted depth exhibit smaller relative degradation, but remain less faithful to the input geometry thus achieving lower overall performance. We also simulate camera perturbations in Tb. 2, where our method and ShapeR are both resilient to

moderate camera noise through use of Plücker embeddings. Other baselines are not evaluated as they do not condition on camera parameters. We additionally report results using 2,048 input points for all methods in Tb. 3. Our method shows stable performance under weaker geometric conditioning while ShapeR saw larger performance drop in single-view compared to Tb. 1.

#### 4.4 Qualitative Results

Fig. 4a and 4b show qualitative results on challenging occlusion scenarios (see Supplement for results without occlusion). Axolotl3D leverages conditioning points to produce faithful reconstructions, retaining details in known regions that other methods may simplify. With our data augmentation, our method learns to exploit the natural symmetry of objects, handling difficult regions—such as the deer in Fig. 4a—that others struggle to reconstruct. In certain examples, baselines produce unwanted structures, whereas our method preserves structural consistency with the provided guidance. Even in failure cases, when our method fails to complete all missing areas, our approach maintains high visual quality in the regions constrained by the conditioning points.

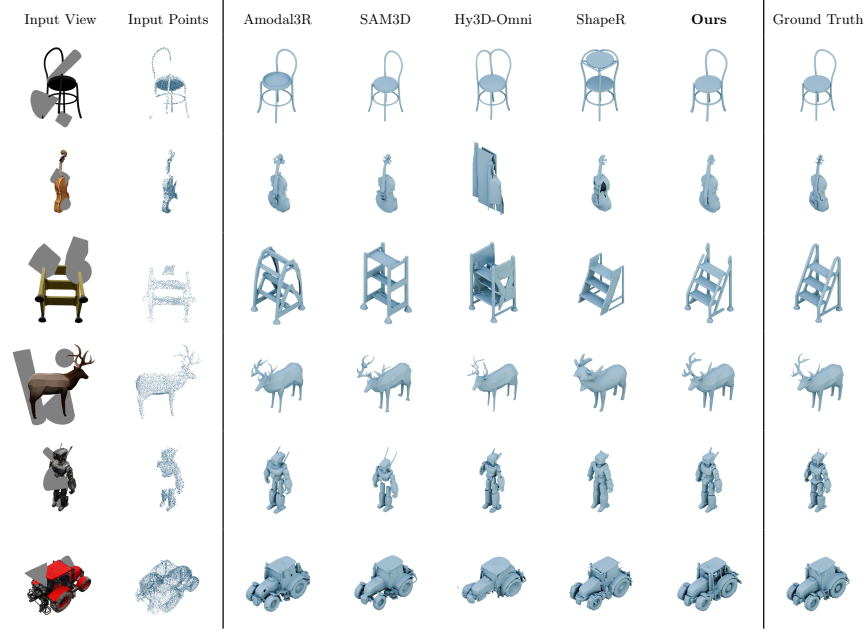
Fig. 5 evaluates robustness under single-view predicted depth. While SAM3D produces high-quality reconstructions, the resulting shapes often deviate from the input point geometry. Hy3D-Omni shows occasional inconsistencies in object orientation, likely due to the absence of explicit camera conditioning. ShapeR is the most sensitive to noise in the predicted depth points, leading to noticeable degradation in mesh quality. In contrast, our method produces reconstructions that are more consistent with the input observations, while remaining robust under moderate depth noise, though performance degrades in cases of severe depth estimation errors.

To stress test our method, we show examples under increasing amounts of occlusion in Fig. 6. Our approach effectively handles severe occlusions by using information from similar unoccluded regions, as shown with the bike wheels. We also include challenging cases such as the horse, where thin structures sometimes fail to fully connect, highlighting areas for improvement.

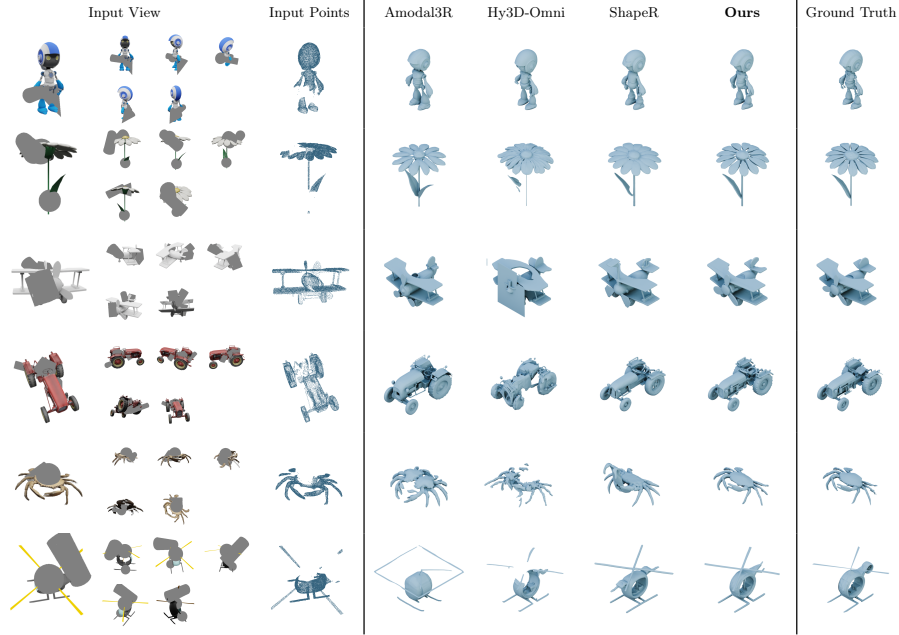
#### 4.5 Ablation Study

We conducted ablations to analyze the contribution of the various components of Axolotl3D and report the results in Tb. 4 and Fig. 7. For these experiments, we evaluate using Toys4k dataset under single- and multi-view occlusion scenarios.

**Effect of Point Condition:** To analysis the importance of geometric conditioning, we remove the point condition along with multi-modal fusion layers, only passing in the output from visual encoding to the cross-attention layers (see §3.1). Removing the point condition leads to the largest drop in quantitative performance and visibly poorer completed objects, indicating that explicit geometric conditioning is crucial for maintaining accuracy in known regions and producing overall more consistent and complete 3D geometry.



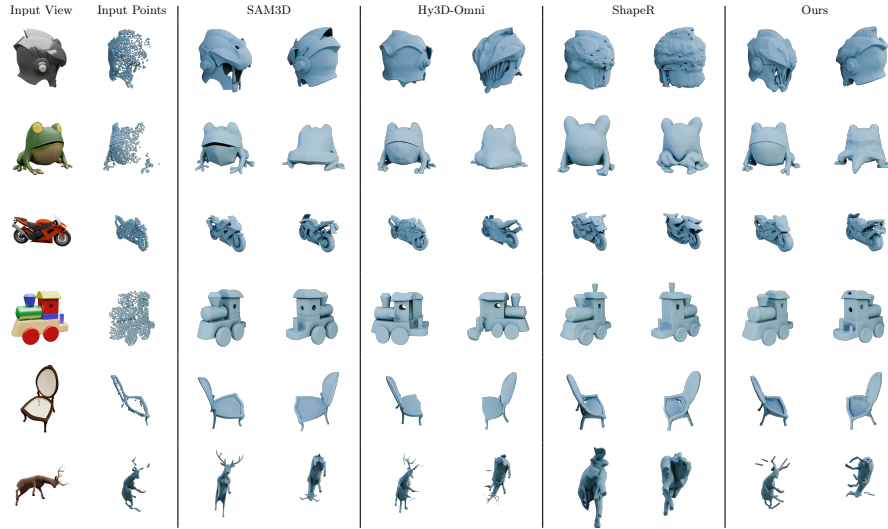
(a) Single-view input results



(b) Multi-view input results

**Fig. 4: Qualitative Comparisons:** Single- and multi-view completion results against baselines. Occluders are shown in grey in the input view. Zoom in for details.





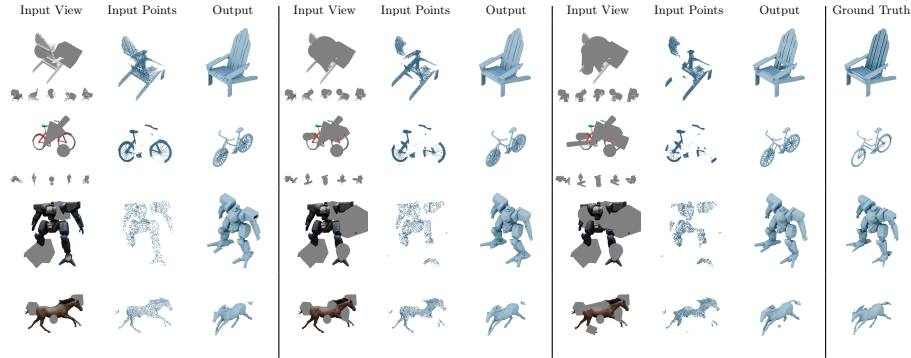
**Fig. 5: Predicted Depth Comparisons:** Unlike SAM3D, our method prioritizes adherence to input points, yielding more faithful completion, but degrades under severe depth prediction errors—a performance trade off. Zoom in for details.

**Effect of Masked Attention:** To assess the contribution of visibility-aware attention, we remove the attention bias from image masks applied to cross-attention layers. Removing the masking causes a small quantitative drop, while visual differences remain minimal, mostly appearing as subtle changes to global geometry such as slight thickness variations. Fine surface details and appearance remain largely unchanged, as these are captured by the unmasked image tokens. This suggests that masking primarily serves as a safeguard against potential interference from occluded regions, influencing global structural reasoning rather than perceptually noticeable features.

**Effect of Camera Condition:** To examine the impact of explicit camera parameterization, we replace camera Plücker embeddings with per-view learnable embeddings to differentiate between the patch tokens from each view. In multi-view settings, metrics remain close to the full model, suggesting the network can largely infer relative camera positions from visual cues when multiple views are available. In single-view scenarios, performance degrades more noticeably, as the model struggles to align 3D points with image features without explicit camera information. This highlights the importance of camera conditioning for reliable object completion under sparse observations.

## 4.6 Applications

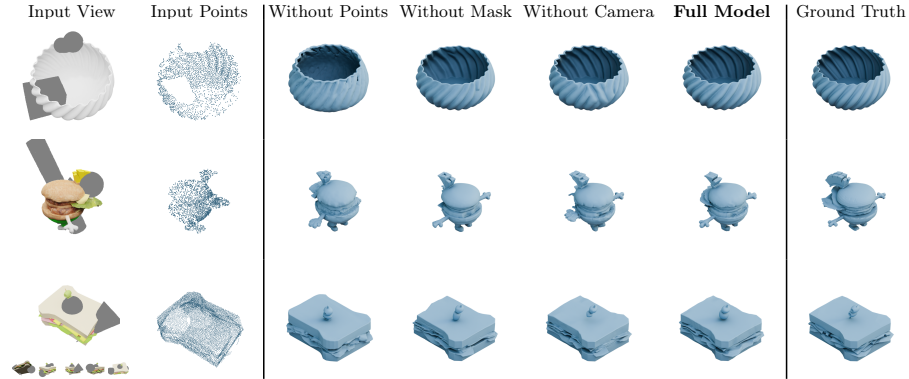
**Shape Editing:** Our method can also be effectively applied for 3D shape editing (Fig.8a). The user selects a view, masks and edits it, e.g. using an image inpainting model like [33]. This forms the single-view condition to our model.



**Fig. 6: Completion Stress Test:** Model performance on challenging single- and multi-view examples with increasing amounts of occluders (3, 5, 7). Zoom in for details.

**Table 4: Ablation Study.** Best and second-best results are highlighted accordingly.

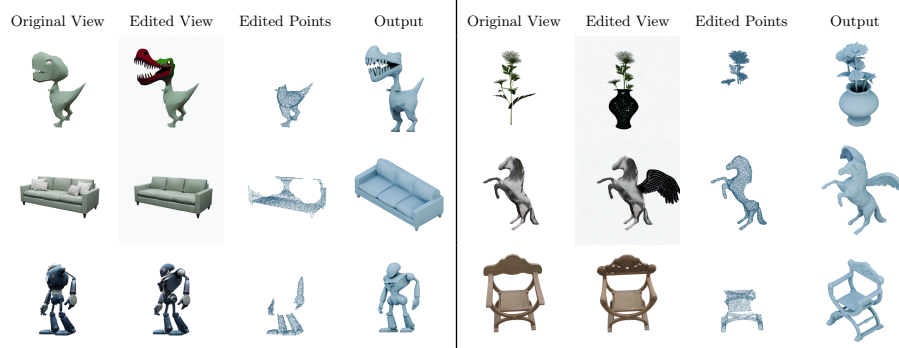
| Variant           | Single-View         |                     |                           | Multi-View          |                     |                           |
|-------------------|---------------------|---------------------|---------------------------|---------------------|---------------------|---------------------------|
|                   | F-score $\uparrow$  | vIoU $\uparrow$     | CD $\downarrow \times 10$ | F-score $\uparrow$  | vIoU $\uparrow$     | CD $\downarrow \times 10$ |
| w/o Points        | 0.8074 $\pm$ 0.1738 | 0.2253 $\pm$ 0.0969 | 0.3589 $\pm$ 0.2993       | 0.9097 $\pm$ 0.0989 | 0.2881 $\pm$ 0.0971 | 0.2231 $\pm$ 0.1344       |
| w/o Mask          | 0.8968 $\pm$ 0.1311 | 0.3225 $\pm$ 0.1068 | 0.2517 $\pm$ 0.2568       | 0.9655 $\pm$ 0.0597 | 0.4151 $\pm$ 0.1192 | 0.1463 $\pm$ 0.0751       |
| w/o Cameras       | 0.9002 $\pm$ 0.1224 | 0.3228 $\pm$ 0.1048 | 0.2509 $\pm$ 0.2414       | 0.9689 $\pm$ 0.0582 | 0.4153 $\pm$ 0.1149 | 0.1462 $\pm$ 0.0878       |
| <b>Full Model</b> | 0.9036 $\pm$ 0.1164 | 0.3262 $\pm$ 0.1039 | 0.2423 $\pm$ 0.2167       | 0.9688 $\pm$ 0.0535 | 0.4185 $\pm$ 0.1123 | 0.1450 $\pm$ 0.0685       |



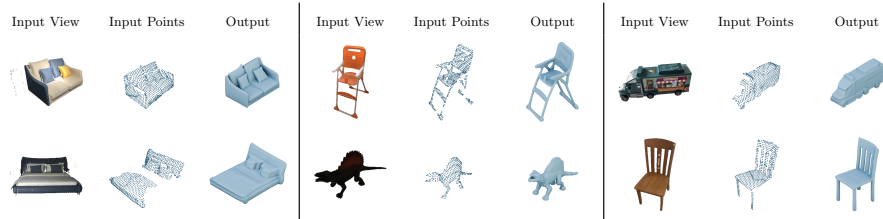
**Fig. 7: Ablation Examples:** Comparing completion results for the full model and ablated variants. Occluders are shown in grey in the input view. Zoom in for details.

To provide conditioning points, we sample the surface of the original shape and discard points that project into the masked region of the edited view. Fig.8a shows effective 3D edits that remain faithful to the unedited parts of the shape.

**Image to 3D:** While our method is trained on points sampled from meshes, it can be combined with approaches like [22] to extend its usability to scenarios where points are not available. For example, to produce 3D shapes from a single image (Fig.8b), we segment the image using SAM 2 [32] and estimate object points and camera parameters using MapAnything [22]. The object points are



(a) **Shape editing** with user edited single-view image through image inpainting.



(b) **Image to 3D** with SAM 2 [32] segmentation and MapAnything [22] outputs.

**Fig. 8: Applications:** Axolotl3D enables precise object edits and accurate single-view generation with completion of missing regions. Zoom in for details.

aligned and normalized, before being passed to our method (see Supplement), resulting in faithful 3D generations from a single view (Fig.8b).

**Simulation:** Accurate and faithful completion is also important for physical applications, such as physics simulation of segmented captured scenes, as shown in [7, 13]. See Supplement for examples of this application.

## 5 Conclusion

We presented Axolotl3D, a unified framework for multi-modal, occlusion-aware 3D generation that completes and edits objects from partial observations. By jointly leveraging images, visibility masks, camera parameters, and partial point clouds, our model enforces geometric consistency while integrating visual and 3D evidence. While previous methods focus on specific tasks in isolation, Axolotl3D provides a single framework capable of handling diverse reconstruction and editing scenarios. Experiments show state-of-the-art performance in sparse-view conditions with and without occlusions. These results highlight its potential for real-world 3D reconstruction and controllable geometry-aware editing. While Axolotl3D is trained to faithfully follow input points, future work could explore balancing this fidelity with robustness to noisy point clouds, as well as supporting more diverse camera positions and textures.

## References

1. Barda, A., Gadelha, M., Kim, V.G., Aigerman, N., Bermano, A.H., Groueix, T.: Instant3dit: Multiview inpainting for fast editing of 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16273–16282 (June 2025)
2. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. Spie (1992)
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
4. Chang, J., Ye, C., Wu, Y., Chen, Y., Zhang, Y., Luo, Z., Li, C., Zhi, Y., Han, X.: Reconviagen: Towards accurate multi-view 3d object reconstruction via generation. In: The Fourteenth International Conference on Learning Representations (2026), <https://openreview.net/forum?id=z0QLeeoEEf>
5. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. arXiv preprint arXiv:2511.16624 (2025)
6. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., Guillaumin, M., Malik, J.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21126–21136 (June 2022)
7. Dagli, R., Xiang, D., Modi, V., Loop, C., Tsang, C.F., Chen, A.H., Hu, A., State, G., Levin, D.I., Shugrina, M.: Vomp: Predicting volumetric mechanical property fields. arXiv preprint arXiv:2510.22975 (2025)
8. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems* **36**, 35799–35813 (2023)
9. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(3), 611–625 (2018). <https://doi.org/10.1109/TPAMI.2017.2658577>
10. Erkoç, Z., Gümeli, C., Wang, C., Nießner, M., Dai, A., Wonka, P., Lee, H.Y., Zhuang, P.: Predictor3d: Fast and precise 3d shape editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 640–649 (June 2025)
11. Feng, H., Liu, J., Tang, J., Wu, G., Chen, B., Lai, J., Wang, G.: Objfiller-3d: Consistent multi-view 3d inpainting via video diffusion models. arXiv preprint arXiv:2508.18271 (2025)
12. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision* **129**(12), 3313–3337 (2021). <https://doi.org/10.1007/s11263-021-01534-z>
13. Fuji Tsang, C., Hu, A., Perel, O., Kolve, C., Shugrina, M.: Artisans: Interactive tools for gaussian splat selection with ai and human in the loop. arXiv preprint arXiv:2602.10173 (2026)
14. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1026–1034 (2015). <https://doi.org/10.1109/ICCV.2015.123>
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020), [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/4c5bcbfec8584af0d967f1ab10179ca4b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcbfec8584af0d967f1ab10179ca4b-Paper.pdf)
17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), <https://openreview.net/forum?id=qw8AKxfYbI>
18. Hsiao, T.F., Ruan, B.K., Liu, Y.L., Shuai, H.H.: Vecset-edit: Unleashing pre-trained lrm for mesh editing from single image. arXiv preprint arXiv:2602.04349 (2026)
19. Huang, Z., Guo, Y.C., An, X., Yang, Y., Li, Y., Zou, Z.X., Liang, D., Liu, X., Cao, Y.P., Sheng, L.: Midi: Multi-instance diffusion for single image to 3d scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 23646–23657 (June 2025)
20. Hunyuan3D, T., Yang, S., Yang, M., Feng, Y., Huang, X., Zhang, S., He, Z., Luo, D., Liu, H., Zhao, Y., et al.: Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. arXiv preprint arXiv:2506.15442 (2025)
21. Hunyuan3D, T., Zhang, B., Guo, C., Liu, H., Yan, H., Shi, H., Huang, J., Yu, J., Li, K., Wang, P., et al.: Hunyuan3d-omni: A unified framework for controllable generation of 3d assets. arXiv preprint arXiv:2509.21245 (2025)
22. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zaus, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S.R., Richardt, C., Ramanan, D., Scherer, S., Kotschieder, P.: MapAnything: Universal feed-forward metric 3D reconstruction. In: International Conference on 3D Vision (3DV). IEEE (2026)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions On Graphics (TOG) **42**(4) (July 2023), <https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/>
24. Khanna, M., Mao, Y., Jiang, H., Haresh, S., Shacklett, B., Batra, D., Clegg, A., Undersander, E., Chang, A.X., Savva, M.: Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16384–16393 (June 2024)
25. Kong, X., Liu, S., Lyu, X., Taher, M., Qi, X., Davison, A.J.: Eschernet: A generative model for scalable view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9503–9513 (June 2024)
26. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Zhao, Y., Peng, S., Guo, H., Zhou, X., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: The Fourteenth International Conference on Learning Representations (2026), <https://openreview.net/forum?id=yirunib818>
27. Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J., Su, H.: One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10072–10083 (June 2024)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019), <https://openreview.net/forum?id=Bkg6RiCqY7>

29. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
30. Ozguroglu, E., Liu, R., Surís, D., Chen, D., Dave, A., Tokmakov, P., Vondrick, C.: pix2gestalt: Amodal segmentation by synthesizing wholes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3931–3940 (June 2024)
31. Qu, Y., Dai, S., Li, X., Wang, Y., Shen, Y., Cao, L., Ji, R.: Deocc-1-to-3: 3d de-occlusion from a single image via self-supervised multi-view diffusion. arXiv preprint arXiv:2506.21544 (2025)
32. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollar, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=Ha6RTeWMd0>
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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 (June 2022)
34. Siddiqui, Y., Frost, D., Aroudj, S., Avetisyan, A., Howard-Jenkins, H., DeTone, D., Moulon, P., Wu, Q., Li, Z., Straub, J., Newcombe, R., Engel, J.: Shaper: Robust conditional 3d shape generation from casual captures. arXiv preprint arXiv:2601.11514 (2026)
35. Sitzmann, V., Rezhikov, S., Freeman, B., Tenenbaum, J., Durand, F.: Light field networks: Neural scene representations with single-evaluation rendering. In: Advances in Neural Information Processing Systems. vol. 34, pp. 19313–19325 (2021), [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/a11ce019e96a4c60832eadd755a17a58-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/a11ce019e96a4c60832eadd755a17a58-Paper.pdf)
36. Stojanov, S., Thai, A., Rehg, J.M.: Using shape to categorize: Low-shot learning with an explicit shape bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1798–1808 (June 2021)
37. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2149–2159 (January 2022)
38. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5294–5306 (June 2025)
39. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2023)
40. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.:  $\pi^3$ : Permutation-equivariant visual geometry learning. In: The Fourteenth International Conference on Learning Representations (2026), <https://openreview.net/forum?id=DTQIjngDta>
41. Wu, T., Zheng, C., Guan, F., Vedaldi, A., Cham, T.J.: Amodal3r: Amodal 3d reconstruction from occluded 2d images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9181–9193 (2025)

42. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., Lin, D., Liu, Z.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 803–814 (June 2023)
43. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., Yang, J.: Native and compact structured latents for 3d generation. Tech report (2025)
44. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21469–21480 (2025)
45. Xu, C., Li, A., Chen, L., Liu, Y., Shi, R., Su, H., Liu, M.: Sparp: Fast 3d object reconstruction and pose estimation from sparse views. In: Proceedings of the European Conference on Computer Vision (ECCV). p. 143–163 (2024)
46. Yan, H., Luo, K., Li, W., Liang, Y., Li, S., Huang, J., Guo, C., Tan, P.: Posemaster: Generating 3d characters in arbitrary poses from a single image. arXiv preprint arXiv:2506.21076 (2025)
47. Yao, K., Zhang, L., Yan, X., Zeng, Y., Zhang, Q., Xu, L., Yang, W., Gu, J., Yu, J.: Cast: Component-aligned 3d scene reconstruction from an rgb image. ACM Transactions On Graphics (TOG) **44**(4) (2025). <https://doi.org/10.1145/3730841>, <https://doi.org/10.1145/3730841>
48. Yu, X., Xu, M., Zhang, Y., Liu, H., Ye, C., Wu, Y., Yan, Z., Zhu, C., Xiong, Z., Liang, T., Chen, G., Cui, S., Han, X.: Mvimngnet: A large-scale dataset of multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9150–9161 (June 2023)
49. Zhang, B.: Vecsetx. <https://github.com/1zb/VecSetX> (2025)
50. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. ACM Transactions On Graphics (TOG) **42**(4), 1–16 (2023)
51. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. ACM Transactions On Graphics (TOG) **43**(4) (2024). <https://doi.org/10.1145/3658146>
52. Zhang, X., Irshad, M.Z., Yezzi, A., Tsai, Y.C., Kira, Z.: Eschernet++: Simultaneous amodal completion and scalable view synthesis through masked fine-tuning and enhanced feed-forward 3d reconstruction. arXiv preprint arXiv:2507.07410 (2025)
53. Zhou, J., Tai, Y.W.: Amodalgen3d: Generative amodal 3d object reconstruction from sparse unposed views. arXiv preprint arXiv:2511.21945 (2025)
54. Zuo, Q., Gu, X., Dong, Y., Zhao, Z., Yuan, W., Qiu, L., Bo, L., Dong, Z.: High-fidelity 3d textured shapes generation by sparse encoding and adversarial decoding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 52–69 (2024). [https://doi.org/10.1007/978-3-031-72684-2\\_4](https://doi.org/10.1007/978-3-031-72684-2_4)