UNIFORGE3D: JOINT 3D RECONSTRUCTION AND GENERATION IN A SHARED ANCHOR-GUIDED SPACE

Anonymous authors
Paper under double-blind review



Figure 1: UniForge3D generates faithful 3D assets from sparse, unposed multi-view images.

ABSTRACT

Generating complete 3D assets from sparse, unposed images requires preserving observed geometry while completing unseen regions. Joint reconstruction and generation offers a way to address both goals: reconstruction recovers the geometry observed in the input images, while generation contributes learned shape priors. However, combining camera-relative reconstruction with canonical-space generation introduces a coordinate mismatch that makes direct geometric interaction more difficult. Moreover, conditioning each 3D location on all views without accounting for their relevance may mix useful geometric evidence with unrelated image content, hindering local detail recovery. We present **UniForge3D**, which jointly models reconstruction and generation in a shared anchor-guided space, resolving their coordinate mismatch. This space follows the anchor view's azimuth while keeping the object upright. The two branches exchange features bidirectionally to predict per-view point maps and a complete sparse voxel structure. We further condition point-map prediction on the generated structure to establish pixel-to-voxel correspondences and select the most informative input view for each voxel. Features from the selected views then guide detailed shape generation, helping preserve observed geometric details. Our method achieves state-of-the-art appearance and geometry results on Toys4K and GSO, generating complete 3D assets with improved fidelity.

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

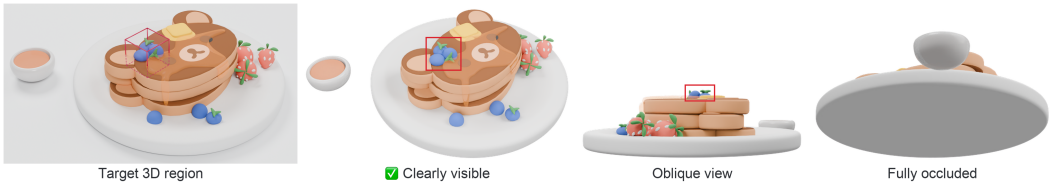

Figure 2: **Different views provide different evidence for the same 3D region.** The 3D box marks three blueberries, which are clearly visible, seen at an oblique angle, or fully occluded across the three views, motivating region-specific view selection.

1 INTRODUCTION

High-quality 3D assets are essential for digital content creation, gaming, augmented and virtual reality, and embodied intelligence. In many practical scenarios, an object is observed through only a few images with unknown camera poses. From these sparse, unposed images, we aim to generate a complete 3D asset that faithfully preserves the geometry and appearance observed in every input view.

Existing multi-view 3D generation methods commonly condition native 3D generators on features extracted from multiple input images. They incorporate multi-view evidence by assigning images to predefined view slots or adaptively fusing features across views (Team, 2025; Li et al., 2026b; Sun et al., 2026; Li et al., 2026a; Zhou et al., 2026). However, image conditioning does not strictly constrain the generated surfaces to match the observed geometry. Generative models may therefore favor plausible shapes under their learned priors over faithfully preserving input-specific structures. This motivates using multi-view reconstruction to provide more direct geometric guidance.

A straightforward approach is to reconstruct geometry from multi-view images and use the resulting implicit features or explicit geometry to condition generation (Chang et al., 2026; Huang et al., 2026; Siddiqui et al., 2026; Hu & Shugrina, 2026; Zhou & Tai, 2025; Liu et al., 2026). However, errors in these upstream reconstruction estimates can propagate into generation, limiting fidelity when the geometric guidance is unreliable.

A further direction explores joint reconstruction and generation. Mix3R (Lin et al., 2026) connects pretrained reconstruction and generation models through a Mixture-of-Transformers architecture, allowing the two tasks to exchange information and improve each other. However, the two branches retain different coordinate representations, requiring an additional predicted similarity transformation to align the reconstructed point maps with the generated geometry.

This coordinate mismatch hinders direct geometric interaction between reconstruction and generation, posing the first challenge. A further challenge lies in recovering local shape details from complex multi-view observations. As illustrated in Fig. 2, the same region may be fully visible with clear details in one view, but difficult to distinguish in another because of an oblique viewing angle or occlusion. Combining all views without accounting for these differences can mix useful evidence with unrelated image content, making it harder to recover details at the corresponding 3D location.

To address these challenges, we propose **UniForge3D**, a framework that jointly models 3D reconstruction and generation in a shared anchor-guided space and uses structure-guided view conditioning to recover local shape details from sparse, unposed multi-view images.

Specifically, **UniForge3D** consists of two main components. First, **Anchor-Guided Joint Reconstruction and Generation** defines a shared space whose horizontal orientation follows the anchor view's azimuth while preserving the object's upright direction, eliminating the coordinate mismatch between the two branches. The branches exchange features bidirectionally to predict per-view point maps and a complete sparse voxel structure in this space. The shared coordinate system allows image observations to be directly mapped onto the generated structure through the point maps. Furthermore, to establish more reliable pixel-to-voxel correspondences, **Structure-Guided View Selection and Conditioning** uses the generated sparse structure to guide point-map prediction. These correspondences identify the most informative input view for each voxel, whose features then guide local shape generation to better preserve the observed geometric details.

Extensive experiments on the Toys4K (Stojanov et al., 2021) and GSO (Downs et al., 2022) datasets demonstrate that **UniForge3D** achieves state-of-the-art performance across all evaluated geometry and appearance metrics, generating complete 3D assets that faithfully preserve the structures and visual details observed in sparse, unposed input views.

2 Related Work

2.1 Single-View 3D Generation

Single-view 3D object generation has progressed rapidly, with existing approaches broadly categorized into methods based on 2D generative priors and native 3D generative methods. Within the former category, one paradigm optimizes 3D representations under supervision distilled from pretrained 2D diffusion models (Poole et al., 2022; Lin et al., 2023; Wang et al., 2023; Melas-Kyriazi et al., 2023; Tang et al., 2023; Qian et al., 2024; Tang et al., 2024b; Liu et al., 2023). The other paradigm employs image or video diffusion models to synthesize multiple views of an object and aggregates the resulting multi-view information to recover 3D representations (Shi et al., 2023; Liu et al., 2024b; Long et al., 2024; Xu et al., 2024a; Tang et al., 2024a; Xu et al., 2024b; Liu et al., 2024a; Li et al., 2024a; Zuo et al., 2024; Wu et al., 2024).

Recent research has increasingly shifted toward native 3D generation, where generative models are trained directly on large-scale 3D data. Driven by advances in 3D latent representations, current approaches primarily follow two paradigms: Vector Set (VecSet) (Zhang et al., 2023; Li et al., 2024b; Zhang et al., 2024; Li et al., 2025b; Hunyuan3D et al., 2025; Li et al., 2025a) and Sparse Voxel (Ren et al., 2024; Xiang et al., 2025; Wu et al., 2026; He et al., 2025; Chen et al., 2026b; Li et al., 2026c; Xiang et al., 2026; Chen et al., 2026a; Ye et al., 2025). VecSet-based methods encode 3D shapes as collections of latent vectors, providing compact representations and effective global modeling, whereas Sparse-Voxel-based methods retain explicit spatial organization and better support local geometric details.

2.2 Multi-View 3D Reconstruction and Generation

Classical multi-view reconstruction combines Structure-from-Motion and multi-view stereo (Schonberger & Frahm, 2016; Furukawa et al., 2010), while learning-based methods improve depth estimation through learned matching (Yao et al., 2018; Gu et al., 2020; Wang et al., 2021). Neural rendering and feed-forward Gaussian prediction further support reconstruction and novel-view synthesis (Mildenhall et al., 2021; Niemeyer et al., 2022; Charatan et al., 2024; Zhang et al., 2025). Recent feed-forward models directly estimate point maps and other geometric quantities from unposed images (Wang et al., 2024; Leroy et al., 2024; Yang et al., 2025; Wang et al., 2025b;a; 2026b; Keetha et al., 2026; Lin et al., 2025; Wang et al., 2026a). However, their reconstructions primarily cover observed surfaces and may remain incomplete in unseen regions.

Multi-view 3D generation follows two main approaches. Training-free methods extend pretrained single-view generators by adaptively selecting or fusing information across views (Li et al., 2026a; Zhou et al., 2026; Qu et al., 2026). Training-based methods learn to incorporate multiple views through feature conditioning or adaptive view routing (Team, 2025; Li et al., 2026b; Sun et al., 2026). These strategies improve the use of multi-view evidence, but image conditioning alone does not ensure that generated surfaces faithfully match the observed geometry.

To combine reconstruction with generative priors, one line of work uses estimated reconstruction features or geometry as generation conditions (Chang et al., 2026; Huang et al., 2026; Siddiqui et al., 2026; Hu & Shugrina, 2026; Zhou & Tai, 2025; Liu et al., 2026). For example, ReconViaGen uses reconstruction features, while RecGen3D additionally incorporates canonicalized point clouds. Such guidance provides geometric and cross-view information, but errors in the reconstruction estimates may propagate into generation. Another direction jointly models reconstruction and generation. Mix3R (Lin et al., 2026) enables bidirectional interaction between the two branches, but retains camera-relative reconstruction and canonical-space generation. It therefore requires learning cross-space alignment alongside both tasks, complicating direct geometric interaction between the two branches. Our method instead places reconstruction and sparse-structure generation in a shared anchor-guided space, giving both branches a consistent geometric reference. We further use structure-guided reconstruction to identify informative input views for detailed shape generation.

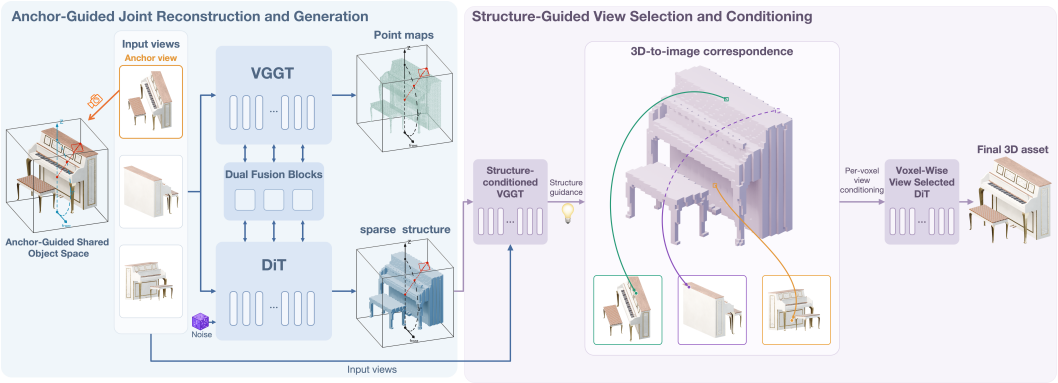

Figure 3: **Overview of UniForge3D. Left:** The reconstruction and generation branches interact bidirectionally to predict multi-view point maps and a complete sparse structure in a shared anchor-guided space. **Right:** Structure-conditioned point maps establish voxel-to-pixel correspondences, enabling the model to select informative views and use their image features to guide detailed shape generation.

3 METHOD

Overview. Given a sparse set of unposed images $\mathcal{I} = \{I_i\}_{i=1}^{N}$ of an object, we aim to generate a complete 3D asset that faithfully preserves the geometry observed in every input view. We designate I_1 as the anchor view and jointly predict multi-view point maps and a complete coarse voxel structure in an anchor-guided shared object space (Section 3.1). The shared coordinate system further allows reconstructed 3D points to be directly associated with generated voxels. We therefore use structure-guided reconstruction to establish explicit pixel-to-voxel correspondences and select the most informative view for each voxel. The selected-view image features then guide detailed shape generation (Section 3.2). The overall pipeline is illustrated in Fig. 3.

3.1 ANCHOR-GUIDED JOINT RECONSTRUCTION AND GENERATION

To achieve high-fidelity multi-view 3D object generation, we seek to make reconstruction and generation mutually beneficial. Directly conditioning generation on reconstructed geometry or features can propagate reconstruction errors. Joint training enables mutual assistance, but their different coordinate systems complicate geometric interaction: feed-forward reconstruction (Wang et al., 2025a) predicts geometry in the first-camera space, whereas native 3D generation (Xiang et al., 2026) operates in a canonical object space. We therefore introduce a shared anchor-guided space that follows the anchor view's azimuth while preserving the canonical vertical direction. Both branches operate in this shared space and exchange features bidirectionally, so that observed geometry guides structure generation and complete-shape context improves reconstruction.

3.1.1 ANCHOR-GUIDED SHARED OBJECT SPACE

We define a shared object space whose front direction follows the azimuth of the anchor view I_1, while its up direction remains aligned with the canonical vertical axis. For any object point $\mathbf{p}^{\text{can}}$ in the centered and scale-normalized canonical space, its shared-space coordinate is

$$\mathbf{p}^{\star} = R_a \mathbf{p}^{\text{can}}, \tag{1}$$

where R_a rotates about the canonical vertical axis so that the anchor's viewing azimuth defines the front direction in the shared space.

We construct the targets of both branches from the transformed geometry. The visible surface points form the per-view point-map targets for reconstruction, while the complete geometry is voxelized to form the sparse-structure target for generation. Consequently, the two branches describe geometry in the same object space and can interact without additional alignment.

3.1.2 BIDIRECTIONAL JOINT MODELING

The two branches contain complementary geometric information. The reconstruction branch describes surfaces observed in the input views, while the generation branch models the complete sparse structure. We connect them through bidirectional interactions at selected layers. We denote the generation tokens by $\mathbf{G}_t^l$ and the reconstruction tokens of view v by $\mathbf{R}_v^l$.

Reconstruction-to-generation. For each view, we spatially align the reconstruction tokens with the DINOv3 (Siméoni et al., 2025) patch features. The reconstruction features predict a residual that is added to the corresponding DINOv3 features, producing a view-specific condition $\mathbf{C}_v^l$. We then concatenate the conditions from all views along the sequence dimension:

$$\mathbf{C}_{\text{mv}}^l = \text{Concat}_v \left[\mathbf{D}_v + \phi^l(\mathbf{R}_v^l) \right]. \tag{2}$$

The generation tokens attend to $\mathbf{C}_{\text{mv}}^l$ through the cross-attention layers of the generation branch. This allows multi-view geometric and semantic evidence to guide sparse-structure generation.

Generation-to-reconstruction. After the corresponding generation blocks, the reconstruction tokens from all views are concatenated along the sequence dimension and attend to the updated generation tokens:

$$\widetilde{\mathbf{R}}^l = \mathbf{R}^l + g^l(t) \odot \mathcal{A}_{g \to r}^l \left(Q = \mathbf{R}^l, K = V = \widetilde{\mathbf{G}}_t^l \right), \tag{3}$$

where $g^l(t)$ gates the injected features according to the noise level. This interaction injects global shape context into the reconstruction branch, helping it estimate the 3D positions of observed pixels.

We jointly optimize the two branches using

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{point}}. \tag{4}$$

Since both branches are supervised using geometry expressed in the same anchor-guided space, the two branches can exchange geometric information without an additional rigid alignment.

3.2 STRUCTURE-GUIDED VIEW SELECTION AND CONDITIONING

Given a generated sparse voxel structure $\mathcal{V}$ and the input images $\mathcal{I}$, this stage generates a fine-grained shape. A surface region may be clearly visible in one view but occluded or difficult to distinguish in another. To identify the observations relevant to each voxel, we perform structure-guided reconstruction in the shared anchor-guided space. The predicted point maps provide per-pixel 3D coordinates in the same coordinate system as the generated voxels, enabling explicit pixel-to-voxel correspondences. These correspondences allow us to select the most informative view for each voxel based on valid pixel support and use its image features to guide detailed shape generation.

3.2.1 STRUCTURE-GUIDED RECONSTRUCTION

We use the generated sparse structure $\mathcal{V}$ as an additional condition for reconstruction, allowing the network to use both image observations and the current generated geometry when estimating per-pixel 3D positions. We encode $\mathcal{V}$ into a sequence of tokens $\mathbf{S}$, which the reconstruction network attends to through cross-attention at selected layers. The network predicts point maps $\hat{\mathbf{P}}_v^{\mathcal{V}} \in \mathbb{R}^{H \times W \times 3}$ and confidence maps $\hat{\rho}_v^{\mathcal{V}} \in \mathbb{R}^{H \times W}$, along with reconstruction features for subsequent image conditioning. The predicted points share the coordinate system of $\mathcal{V}$, enabling direct association between image pixels and generated voxels.

3.2.2 VOXEL-WISE VIEW SELECTION AND CONDITIONING

We use the predicted point maps to establish pixel-to-voxel correspondences directly in the shared anchor-guided space. After confidence filtering, we quantize the predicted points onto the voxel grid. For each voxel centered at x_j and each view v, we identify the retained pixels mapped to that voxel. We define the support score $s_v(x_j)$ as the number of pixels in the largest connected component on the image. A larger score suggests that the view provides more visual information about the region represented by the voxel. For each voxel, we select

$$v_j^\star = \arg\max_v s_v(x_j). \tag{5}$$

Voxels without direct support inherit a view from a directly supported voxel in their 26-neighborhood; if none exists, they use the anchor view.

For each view, an adapter aligns the Structure-Guided Reconstruction features with the DINOv3 (Siméoni et al., 2025) patch tokens. The aligned features are added to the corresponding DINOv3 tokens to form the view-specific feature sequence. The shape denoiser interleaves two types of cross-attention: **global multi-view attention**, in which every voxel token attends to the features of all views concatenated along the sequence dimension, and **voxel-specific selected-view attention**, in which each voxel token attends only to its selected view $v_j^\star$. The former provides global context for shape completion, while the latter supplies reliable local evidence.

4 Experiments

4.1 Experimental Setup

Data Preparation. Our model is trained on a subset of the TRELLIS-500K dataset (Xiang et al., 2025) containing 210K objects. For each object, we sample two random azimuthal orientations and render six candidate anchor views for each orientation. We additionally render 48 views from randomly sampled viewpoints to provide additional input views for multi-view training. We randomly select 500 objects from Toys4K (Stojanov et al., 2021) and 500 from Google Scanned Objects (GSO) (Downs et al., 2022), for a total of 1,000 evaluation objects. All test objects are held out from training. For each test object, we render four images from randomly selected viewpoints. All methods receive the same four unposed RGB images.

Evaluation Metrics. We evaluate appearance quality using PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). Chamfer Distance (CD) (Fan et al., 2017) is computed as the sum of the bidirectional mean nearest-neighbor distances between points sampled from the reconstructed and ground-truth surfaces. We compute F-score using a distance threshold of 0.01. All appearance and geometry metrics are computed after aligning each reconstructed mesh with its ground-truth mesh using a GICP-based Sim(3) (Segal et al., 2009) transformation. For the ablation on anchor-guided joint modeling, each predicted sparse structure is aligned with the ground truth before evaluating the 64^3 binary surface-occupancy grids using IoU and CD. For point-map reconstruction, predictions from all input views are jointly aligned with the visible ground-truth position maps using a shared object-level Umeyama Sim(3) transformation (Umeyama, 1991). Accuracy measures the mean distance from predicted to ground-truth points, Completeness measures the mean distance in the reverse direction, and Overall is their arithmetic mean. For the view-conditioning ablation, PSNR, SSIM, and LPIPS are computed on rendered normal maps.

Training Details. All models are trained on 32 NVIDIA A800 GPUs. For Anchor-Guided Joint Reconstruction and Generation, we initialize the generation and reconstruction branches from pretrained TRELLIS.2 (Xiang et al., 2026) and VGGT (Wang et al., 2025a), respectively, and fine-tune them using LoRA (Hu et al., 2021) with ranks 64 and 32, while fully training the new interaction modules. We train for 60K iterations with a global batch size of 128 and a learning rate of 5×10^{-5}. For Structure-Guided Reconstruction, we fully fine-tune pretrained VGGT for 40K iterations, then freeze it during subsequent generator training. For coarse-to-fine shape generation, we fine-tune pretrained TRELLIS.2 denoisers using rank-128 LoRA and fully train the new modules. Each model is trained for 60K iterations with a global batch size of 64 and a learning rate of 5×10^{-5}. The texture generator uses the same training settings as the detailed shape generator.

4.2 Comparison with State-of-the-Art Methods

Quantitative Results. We compare our method with Hunyuan3D-2MV (Team, 2025), Mix3R (Lin et al., 2026), MV-SAM3D (Li et al., 2026a), ReconViaGen (Chang et al., 2026), and Stream3D + TRELLIS.2 (Zhou et al., 2026). As shown in Table 1, our method achieves the best performance on both Toys4K (Stojanov et al., 2021) and GSO (Downs et al., 2022) across all appearance and geometry metrics. The improvements in CD and F-score demonstrate more accurate geometry, while the gains in PSNR, SSIM, and LPIPS indicate better appearance preservation under novel viewpoints.

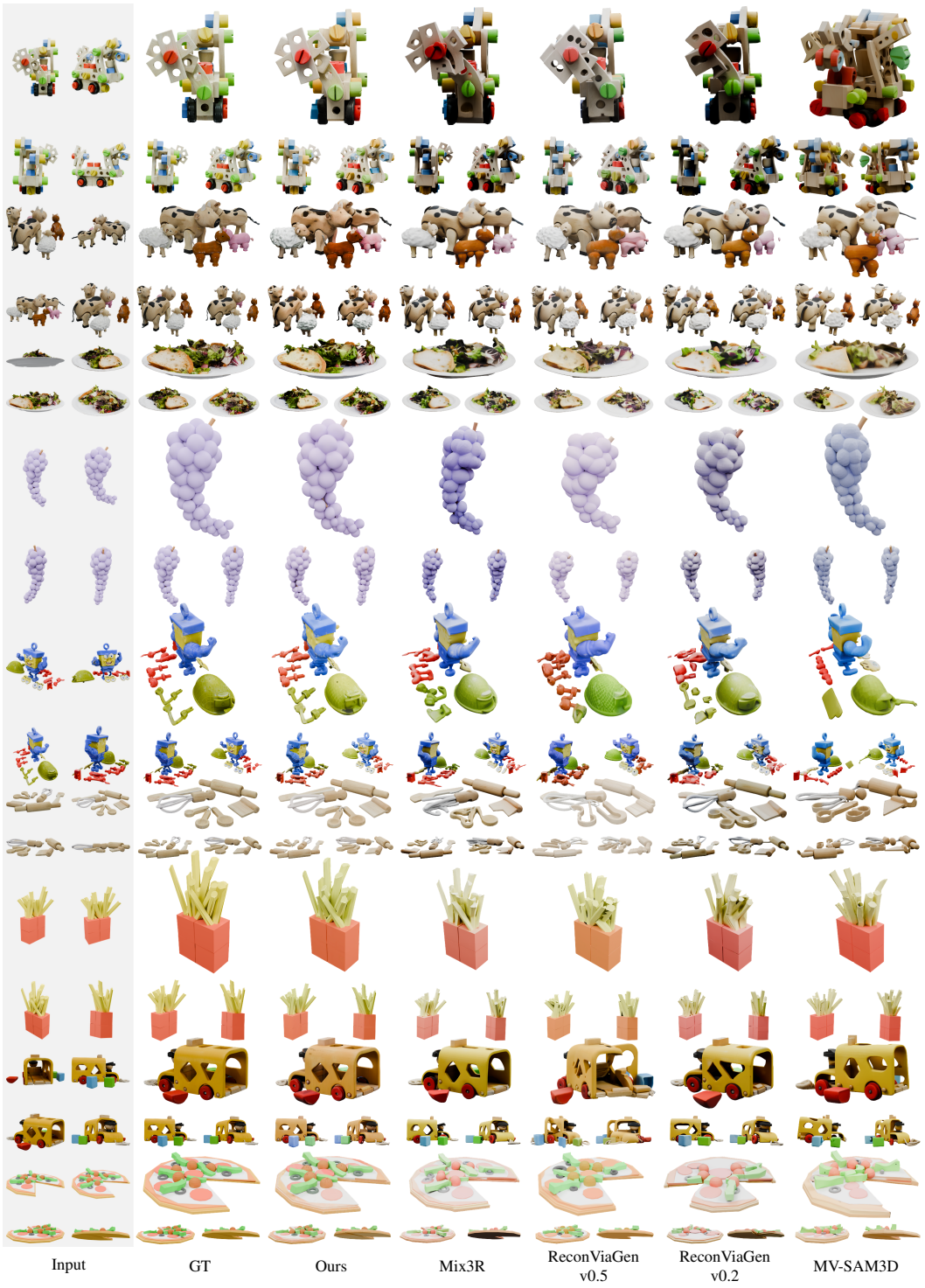

Figure 4: Qualitative comparisons on Toys4K and GSO. Each row shows four input views, ground-truth renderings, and results from different methods. To enable direct comparison, each generated mesh is geometrically aligned with the ground-truth asset and rendered using the corresponding input-view cameras. Our method better preserves the observed appearance and local structures while producing complete and consistent geometry.



Figure 5: Qualitative comparisons on real-world sparse, unposed images. All results are aligned and rendered from common viewpoints for visual comparison.

Table 1: Quantitative comparison on the Toys4K (Stojanov et al., 2021) and GSO (Downs et al., 2022) datasets. Best results are shown in bold and second-best results are underlined. PSNR, SSIM, and LPIPS are computed on RGB renderings.

Dataset	**Toys4K (Stojanov et al., 2021)**					**GSO (Downs et al., 2022)**				
Metric	PSNR↑	SSIM↑	LPIPS↓	CD ($\times 10^{-2}$)↓	F-score↑	PSNR↑	SSIM↑	LPIPS↓	CD ($\times 10^{-2}$)↓	F-score↑
Stream3D + TRELLIS.2 (Zhou et al., 2026)	20.955	0.9351	0.0789	6.3608	0.5003	19.576	0.9119	0.0990	6.2441	0.4700
Hunyuan3D-2MV (Team, 2025)	22.726	0.9470	0.0576	2.9175	0.6723	20.526	0.9206	0.0780	3.2713	0.6139
MV-SAM3D (Li et al., 2026a)	23.071	0.9469	0.0590	2.7147	0.6830	20.653	0.9222	0.0791	2.7447	0.6517
ReconViaGen v0.2 (Chang et al., 2026)	22.713	0.9451	0.0548	2.6155	0.6786	20.498	0.9196	0.0738	2.5905	0.6697
ReconViaGen v0.5 (Chang et al., 2026)	<u>25.020</u>	<u>0.9553</u>	0.0519	2.7007	0.6840	<u>22.774</u>	<u>0.9313</u>	0.0731	2.7002	0.6489
Mix3R (Lin et al., 2026)	23.907	0.9498	<u>0.0506</u>	<u>2.3445</u>	<u>0.7343</u>	21.040	0.9239	<u>0.0688</u>	<u>2.3202</u>	<u>0.7174</u>
Ours	**27.428**	**0.9616**	**0.0401**	**1.4196**	**0.8899**	**24.826**	**0.9363**	**0.0591**	**1.4047**	**0.8897**

Qualitative Results. Fig. 4 and Fig. 5 compare results on synthetic benchmarks and real-world images, respectively. Our method better preserves observed geometry and appearance, while competing methods often omit parts or distort local structures. These results demonstrate its fidelity to the input observations.

4.3 Ablation Study

Ablation on Anchor-Guided Joint Modeling. Table 2 compares different conditioning schemes and coordinate spaces for sparse-structure generation. *DINO Conditioning* concatenates per-view DINOv3 (Siméoni et al., 2025) features along the sequence dimension, with separate embeddings for anchor and non-anchor views. *DINO–VGGT Conditioning* additionally uses VGGT (Wang et al., 2025a), fine-tuned for point-map prediction in our anchor-guided shared space and then frozen. An MLP adapter with a zero-initialized output projection fuses spatially aligned features from four VGGT layers into a residual, which is added to DINOv3 features before concatenating views. Its consistent gains over DINO Conditioning show that reconstruction features complement DINOv3 with geometric cues and cross-view relationships. Our full model, *Anchor-Guided Joint Training*, further jointly optimizes the reconstruction and generation branches through bidirectional feature interaction. It achieves the best results across all metrics, supporting the benefit of joint optimization with bidirectional feature interaction.

We further compare coordinate-space choices for joint training. In *Canonical-Space Joint Training*, both branches operate in the canonical object space. In *Cross-Space Joint Training*, generation operates in the canonical object space while reconstruction remains in VGGT's first-camera space. Both variants perform worse than our anchor-guided formulation, highlighting the benefit of a shared coordinate space whose orientation is tied to the input views.

Table 2: Ablation of conditioning schemes and coordinate spaces for sparse-structure generation on GSO (Downs et al., 2022).

Method	VGGT Space	DiT Space	IoU↑	CD ($\times 10^{-2}$)↓
DINO Conditioning	N/A	Anchor-Guided Space	0.4261	2.0019
DINO–VGGT Conditioning	Anchor-Guided Space	Anchor-Guided Space	<u>0.5533</u>	<u>1.1710</u>
Canonical-Space Joint Training	Canonical Space	Canonical Space	0.4601	1.5825
Cross-Space Joint Training	First-Camera Space	Canonical Space	0.4563	1.6291
Anchor-Guided Joint Training (Ours)	Anchor-Guided Space	Anchor-Guided Space	**0.5827**	**1.0563**

Table 3: Point-map reconstruction results on the GSO dataset (Downs et al., 2022). Distances are reported in normalized object coordinates ($\times 10^{-3}$).

Method	Acc.↓	Comp.↓	Overall↓
Image-Only VGGT	5.025	4.667	4.846
Jointly Trained VGGT	<u>4.053</u>	<u>4.250</u>	<u>4.152</u>
Structure-Conditioned VGGT	**3.539**	**3.086**	**3.313**

Table 4: Ablation of view-conditioning strategies for detailed shape generation on GSO (Downs et al., 2022). Both variants take the same predicted 64^3 sparse structure as input. PSNR, SSIM, and LPIPS are computed on rendered normal maps.

Method	PSNR↑	SSIM↑	LPIPS↓	CD ($\times 10^{-2}$)↓	F-score↑
Full Multi-View Conditioning	26.892	0.9649	0.0401	1.4739	0.8758
Voxel-Wise View Selection and Conditioning (Ours)	**27.207**	**0.9657**	**0.0388**	**1.4047**	**0.8897**

Ablation on Reconstruction Modeling. We ablate the reconstruction design using three VGGT variants. *Image-Only VGGT* is independently fine-tuned end-to-end using only multi-view images. *Jointly Trained VGGT* is jointly optimized with the sparse-structure generator through bidirectional feature interaction in the first stage. *Structure-Conditioned VGGT* further introduces the generated sparse structure as an explicit condition for point-map prediction in the second stage. As shown in Table 3, *Jointly Trained VGGT* outperforms *Image-Only VGGT*, supporting the benefit of interaction with the generation branch for reconstruction. *Structure-Conditioned VGGT* produces the most accurate and complete point maps, demonstrating the benefit of explicit sparse-structure conditioning for reconstruction.

Ablation on Voxel-Wise View Selection and Conditioning. Using the same predicted sparse structure, we compare two conditioning strategies for detailed shape generation. In **Full Multi-View Conditioning**, the adapter-aligned reconstruction features are added to the corresponding DINOv3 patch tokens for each view. The resulting feature sequences from all views are concatenated along the sequence dimension and used to condition every denoiser layer. Our **Voxel-Wise View Selection and Conditioning** alternates full multi-view layers with voxel-specific selected-view layers. As shown in Table 4, our design improves geometric reconstruction performance. These results show our voxel-wise view selection and conditioning strategy improves the fidelity of generated shapes.

5 Limitation and Conclusion

We presented UniForge3D for joint 3D reconstruction and generation from sparse, unposed images. The two branches exchange features in a shared anchor-guided space to predict per-view point maps and a complete sparse structure. The generated structure then guides reconstruction to establish pixel-to-voxel correspondences and select informative views for detailed shape generation. Quantitative results on Toys4K and GSO (Table 1), qualitative comparisons in Figs. 4 and 5, and the supplementary video demonstrate that UniForge3D generates complete 3D assets with faithful geometry and appearance and generalizes well to real-world images.

Limitation. Our model is trained on synthetic renderings, and severe occlusion, strong reflections, and extreme lighting remain challenging. Future work will explore training with simulations of these challenging conditions to improve real-world robustness.

AI USE STATEMENT

We used generative AI tools to assist with translation, drafting and language editing of selected manuscript passages, and checking submission requirements. The authors take responsibility for the accuracy and integrity of the final manuscript, including text prepared with AI assistance.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the proposed method in Section 3 and summarize the training and testing data, rendering procedure, and training configurations in Section 4.1. Appendix A.1 provides further details on the training objectives, parameter adaptation, network architecture, and feature interaction mechanisms. We also provide a supplementary video demonstrating generation results. We plan to publicly release the source code to support reproduction of the reported results.

ETHICS STATEMENT

We use publicly available 3D asset datasets for training and evaluation: a subset of TRELLIS-500K for training, and Toys4K and Google Scanned Objects (GSO) for evaluation. Our method reconstructs object geometry and appearance from input images and can reproduce existing product designs or characters. Before sharing or selling generated models, users should check the licenses of the original assets and obtain permission when required.

REFERENCES

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*, volume 2026, pp. 90387–90401, 2026.

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

Wenyue Chen, Wenjue Chen, Peng Li, Qinghe Wang, Xu Jia, Heliang Zheng, Rongfei Jia, Yuan Liu, and Ronggang Wang. Know3d: Prompting 3d generation with knowledge from vision-language models. In *European Conference on Computer Vision*, pp. 555–573. Springer, 2026a.

Wenyue Chen, Peng Li, Wangguandong Zheng, Chengfeng Zhao, Mengfei Li, Yaolong Zhu, Zhiyang Dou, Ronggang Wang, and Yuan Liu. Synchuman: Synchronizing 2d and 3d generative models for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 38:119488–119514, 2026b.

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

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

Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 1434–1441. IEEE, 2010.

Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2492–2501. IEEE, 2020.

Xianglong He, Zi-Xin Zou, Chia-Hao Chen, Yuan-Chen Guo, Ding Liang, Chun Yuan, Wanli Ouyang, Yan-Pei Cao, and Yangguang Li. Sparseflex: High-resolution and arbitrary-topology 3d shape modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14822–14833. IEEE, 2025.

Anita Hu and Maria Shugrina. Axolotl3d: A unified framework for faithful 3d shape completion. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Zhisheng Huang, Jiahao Chen, Cheng Lin, Chenyu Hu, Hanzhuo Huang, Zhengming Yu, Mengfei Li, Yuheng Liu, Zekai Gu, Zibo Zhao, Yuan Liu, Xin Li, and Wenping Wang. Recgen3d: Reconstruction-guided 3d generation in a shared canonical space. In *SIGGRAPH Asia 2026 Conference Papers*, 2026.

Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, et al. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-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)*. IEEE, 2026.

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

Baicheng Li, Dong Wu, Jun Li, Shunkai Zhou, Zecui Zeng, Lusong Li, and Hongbin Zha. Mv-sam3d: Adaptive multi-view fusion for layout-aware 3d generation. *arXiv preprint arXiv:2603.11633*, 2026a.

Dong-Yang Li, Wang Zhao, Yuxin Chen, Wenbo Hu, Meng-Hao Guo, Fang-Lue Zhang, Ying Shan, and Shi-Min Hu. Pixal3d: Pixel-aligned 3d generation from images. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026b.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *Advances in Neural Information Processing Systems*, 37:55975–56000, 2024a.

Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024b.

Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025a.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Zhihao Li, Yufei Wang, Heliang Zheng, Yihao Luo, and Bihan Wen. Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. *Advances in Neural Information Processing Systems*, 38:118582–118600, 2026c.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300–309. IEEE, 2023.

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

Siyou Lin, Zhou Xue, Hongwen Zhang, Liang An, Dongping Li, Shaohui Jiao, and Yebin Liu. Mix3r: Mixing feed-forward reconstruction and generative 3d priors for joint multi-view aligned 3d reconstruction and pose estimation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026.

Jiarui Liu, Heng Li, Weiyu Li, Keng Deng, Junyuan Deng, Zheng Zhongxing, Junyu Huang, Jiahao Chang, Xiaoguang Han, and Ping Tan. Reconplusgen: Injecting reconstruction prior into multi-view 3d generation through noise inversion and modulation. *arXiv preprint arXiv:2609.11129*, 2026.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024a.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9264–9275. IEEE, 2023.

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*, volume 2024, pp. 27676–27697, 2024b.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9970–9980. IEEE, 2024.

Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion 360 reconstruction of any object from a single image. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8446–8455. IEEE, 2023.

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.

Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5470–5480. IEEE, 2022.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *International Conference on Learning Representations*, volume 2024, pp. 48142–48159, 2024.

Zefan Qu, Zhenwei Wang, Gerhard Petrus Hancke, and Rynson WH Lau. Umi3d: Robust 3d generation on unconstrained multi-image inputs via simultaneous focus cross-attention routing. *arXiv preprint arXiv:2607.24298*, 2026.

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

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Aleksandr Segal, Dirk Haehnel, Sebastian Thrun, et al. Generalized-icp. In *Robotics: science and systems*, volume 2, pp. 435. Seattle, WA, 2009.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

Yawar Siddiqui, Duncan Frost, Samir Aroudj, Armen Avetisyan, Henry Howard-Jenkins, Daniel DeTone, Pierre Moulon, Qirui Wu, Zhengqin Li, Julian Straub, et al. Shaper: Robust conditional 3d shape generation from casual captures. *arXiv preprint arXiv:2601.11514*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Stefan Stojanov, Anh Thai, and James M Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1798–1808. IEEE, 2021.

Hanxiao Sun, Mingxin Yang, Shuhui Yang, Zebin He, Xintong Han, Hongbo Fu, Chunchao Guo, and Wenhan Luo. Roar-3d: Routing arbitrary views for high-fidelity 3d generation. *arXiv preprint arXiv:2605.21121*, 2026.

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*, pp. 1–18. Springer, 2024a.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *International Conference on Learning Representations*, volume 2024, pp. 33879–33896, 2024b.

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22762–22772. IEEE, 2023.

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

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

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14189–14198. IEEE, 2021.

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

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

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

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*, 2026b.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37:125116–125141, 2024.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, Yao Yao, et al. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *Advances in Neural Information Processing Systems*, 38:170778–170804, 2026.

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

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14419–14429, 2026.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024a.

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*, pp. 1–20. Springer, 2024b.

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

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European conference on computer vision*, pp. 785–801. Springer, 2018.

Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 01–12. IEEE, 2025.

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

Chuanrui Zhang, Yingshuang Zou, Zhuoling Li, Minmin Yi, and Haoqian Wang. Transplat: Generalizable 3d gaussian splatting from sparse multi-view images with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9869–9877, 2025.

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

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

Junwei Zhou and Yu-Wing Tai. Gena3d: Generative amodal 3d modeling by bridging 2d priors and 3d coherence. *arXiv preprint arXiv:2511.21945*, 2025.

Kaichen Zhou, Zeyang Bai, Xinhai Chang, Mengyu Wang, Paul Liang, and Fangneng Zhan. Stream3d: Sequential multi-view 3d generation via evidential memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026.

Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Weihao Yuan, Rui Peng, Siyu Zhu, Liefeng Bo, Zilong Dong, Qixing Huang, et al. Videomv: Consistent multi-view generation based on large video generative model. 2024.

A Appendix

A.1 More Details

A.1.1 Training Details

Anchor-Guided Joint Reconstruction and Generation. We train the joint model on 32 NVIDIA A800 GPUs with a global batch size of 128. The generation and reconstruction branches are built upon pretrained TRELLIS.2 (Xiang et al., 2026) and VGGT (Wang et al., 2025a) models, respectively, and are adapted using LoRA (Hu et al., 2021) with ranks 64 and 32. We fully train the newly introduced interaction modules, with zero-initialized condition-residual outputs and time-gating projections, and keep DINOv3 (Siméoni et al., 2025) frozen. We use a peak learning rate of 5×10^{-5}. To enable classifier-free guidance, we randomly drop the image condition with probability 0.1; the same strategy is used in subsequent generative stages.

The training objective combines latent flow matching with point-map supervision:

$$\begin{aligned} \mathcal{L}_{\text{joint}} &= \mathcal{L}_{\text{flow}}^{\text{structure}} + \mathcal{L}_{\text{point}}, \\ \mathcal{L}_{\text{point}} &= \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{normal}}. \end{aligned} \tag{6}$$

Here, $\mathcal{L}_{\text{flow}}$ is the mean squared error between the predicted and target latent velocities. For a valid foreground pixel, let $e_i = \|\widehat{P}_i - P_i\|_2$ denote the point error and $\widehat{\rho}_i$ the predicted confidence. The regression and confidence terms are

$$\mathcal{L}_{\text{reg}} = \mathcal{R}(\{e_i\}), \qquad \mathcal{L}_{\text{conf}} = \mathcal{R}\big(\{\widehat{\rho}_i e_i - 0.2 \log \widehat{\rho}_i\}\big), \tag{7}$$

where $\mathcal{R}$ denotes robust averaging with a 98% quantile filter. $\mathcal{L}_{\text{normal}}$ measures the mean cosine error between point-map normals at three scales.

Structure-Guided Reconstruction. We train this network on 32 NVIDIA A800 GPUs, using dynamic batches of up to 24 images per GPU. We initialize the model from pretrained VGGT and fully fine-tune it with a peak learning rate of 5×10^{-5}. The training objective is

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{normal}}. \tag{8}$$

After training, this network is frozen to provide point maps and reconstruction features for shape and texture generation.

Structure-Guided Shape Generation. We train the 512 and 1024 shape generators separately on 32 NVIDIA A800 GPUs with a global batch size of 64. Both models are initialized from pretrained TRELLIS.2 (Xiang et al., 2026) shape generators, with rank-128 LoRA fine-tuning of the denoisers and full optimization of the feature adapters, while DINOv3 and the Structure-Guided Reconstruction network remain frozen. Both stages use a peak learning rate of 5×10^{-5}. Both stages use latent flow matching on precomputed shape latents, with $\mathcal{L}_{\text{shape}} = \mathcal{L}_{\text{flow}}^{\text{shape}}$.

Shape-Conditioned Texture Generation. We fine-tune the texture generator on 32 NVIDIA A800 GPUs with a global batch size of 64 and a peak learning rate of 5×10^{-5}. We initialize it from a previously trained TRELLIS.2-based texture model and optimize rank-128 denoiser LoRA parameters together with the full feature adapter. The reconstruction network, DINOv3, and latent autoencoders remain frozen. During training, we keep the shape latent fixed and apply flow matching to the texture latent.

All stages use AdamW with cosine learning-rate decay. Each training sample contains four views. In joint training, point-map supervision is restricted to samples retaining the image condition.

A.1.2 Detailed Network Structure

Anchor-Guided Joint Reconstruction and Generation. The joint model comprises a VGGT reconstruction branch and a DiT generation branch. They exchange features bidirectionally in the shared anchor-guided space to jointly predict per-view point maps and a complete sparse structure.

(1) Reconstruction-to-generation. For each view v, we concatenate the frame-wise and cross-view VGGT features into F_v, map them through an adapter ϕ, and add the result to the spatially corresponding DINOv3 image features D_v. We then concatenate all views along the sequence dimension:

$$C_v = D_v + \phi(F_v), \qquad C_{\mathrm{mv}} = \mathrm{Concat}_v[C_v]. \tag{9}$$

DiT uses generation features as queries and C_{mv} as keys and values in cross-attention, allowing generation to use the observed geometry provided by the reconstruction branch.

(2) Generation-to-reconstruction. After updating the generation features, we concatenate the global VGGT features across views into R and use them as queries, with generation features G as keys and values:

$$R' = R + \mathrm{CA}_t(R, G, G). \tag{10}$$

Here, CA_t denotes cross-attention with time-conditioned modulation and gating, allowing the reconstruction branch to receive complete-shape information according to the current denoising state. The updated R' continues through reconstruction. After repeated bidirectional interactions, the two branches output point maps and a sparse structure through the point head and structure decoder, respectively.

Structure-Guided Reconstruction. We further use the generated sparse structure to condition reconstruction. Each voxel center x_j is encoded with Fourier positional encoding and an MLP to obtain a structure feature S_j. Reconstruction features then query these structure features through cross-attention:

$$S_j = \mathrm{MLP}(\mathrm{PE}(x_j)), \qquad R' = R + \mathrm{CA}(R, S, S). \tag{11}$$

Here, S is the sequence of all structure features. We insert this operation into the VGGT backbone and point decoder so that point-map prediction uses both image observations and the complete structure. The network outputs point maps in the shared object space, confidence maps, and multi-level reconstruction features for subsequent generation.

Structure-Guided Shape Generation. Given a sparse structure and multi-view input images, we generate a detailed mesh through Structure-Guided Reconstruction and voxel-wise view conditioning. The sparse structure defines the spatial locations of the shape latent and also serves as a geometric condition for the reconstruction network. Conditioned on this structure, the reconstruction network predicts per-view point maps in the shared object space, confidence maps, and multi-level reconstruction features. These outputs provide both geometric correspondences for view selection and image-aligned features for conditioning shape generation.

For each input view v, we fuse the multi-level reconstruction features with DINOv3 image features to construct the generation condition:

$$C_v = D_v + \psi\big(\mathrm{Concat}_l\big[F_v^l\big]\big)\,, \tag{12}$$

where F_v^l denotes the reconstruction features of view v at feature level l, D_v denotes the corresponding DINOv3 image features, and ψ is a feature adapter that performs multi-level feature mapping, fusion, and spatial alignment. The resulting condition combines reconstruction features informed by the sparse structure with visual features extracted from the input image.

We use the predicted point maps to establish pixel-to-voxel correspondences in the shared anchor-guided space. For each view, we retain the highest-confidence 95% of valid foreground predictions and quantize the predicted points onto the voxel grid.

For a voxel centered at x_j, we group the retained pixels assigned to it in each view v and define the support score $s_v(x_j)$ as the number of pixels in their largest connected component. For voxels with direct support, we select

$$v_j^\star = \arg\max_v s_v(x_j). \tag{13}$$

Ties favor the anchor view, followed by higher mean confidence within the selected component.

A voxel without direct support inherits the selected view of the most strongly supported voxel in its 26-neighborhood. If the neighborhood also provides no support, we use the anchor view, which is the first input view.

During shape generation, DiT alternates between multi-view attention and selected-view attention. Multi-view attention accesses the feature sequences of all input views, providing global context for shape generation. Selected-view attention allows each voxel to access the complete feature sequence $C_{v_j^*}$ of its assigned view. Thus, the selected-view condition is specific to each voxel while retaining the full visual context of the corresponding image. By alternating these two conditioning schemes, the model combines global multi-view information with voxel-associated visual evidence to recover geometric details. Under these conditions, DiT progressively updates the noisy shape latent at the sparse locations defined by the structure. The resulting latent is finally decoded into a mesh by the shape decoder.

Shape-Conditioned Texture Generation. We encode the final mesh into a shape latent and concatenate it with the noisy texture latent at the same sparse locations as input to the texture DiT. The texture model follows the Structure-Guided Reconstruction and view-selection scheme of the shape stage, with an independent adapter for constructing image conditions. During generation, the shape latent remains fixed while the texture latent is progressively updated. The resulting latent is decoded into base color, metallic, roughness, and opacity attributes and baked into UV textures.

A.2 More Results

A.2.1 Qualitative Ablation of Structure Conditioning

Fig. 6 presents additional qualitative comparisons of different VGGT conditioning strategies. Each row corresponds to one object. From left to right, we show four input views, point maps predicted by image-only VGGT, jointly trained VGGT, and structure-conditioned VGGT (ours), followed by the ground-truth visible point maps. Joint training improves geometric consistency over the image-only baseline. Furthermore, explicitly conditioning VGGT on the generated sparse 3D structure produces more complete and coherent geometry.

A.2.2 Qualitative Ablation of View Selection and Conditioning

Fig. 7 provides qualitative results complementing Table 4. Using the same predicted sparse structure, we compare full multi-view conditioning with our alternating full multi-view and selected-view conditioning. The highlighted regions show that our method better preserves local geometric details, demonstrating the benefit of introducing view-specific visual evidence.

A.2.3 Additional Qualitative Results on Rendered Assets

Fig. 8 presents additional results on rendered 3D assets. For comparison with the input observations, we geometrically align each generated mesh with the ground-truth asset and render it from the input cameras. These examples illustrate how our method preserves the observed geometry and appearance across different objects.

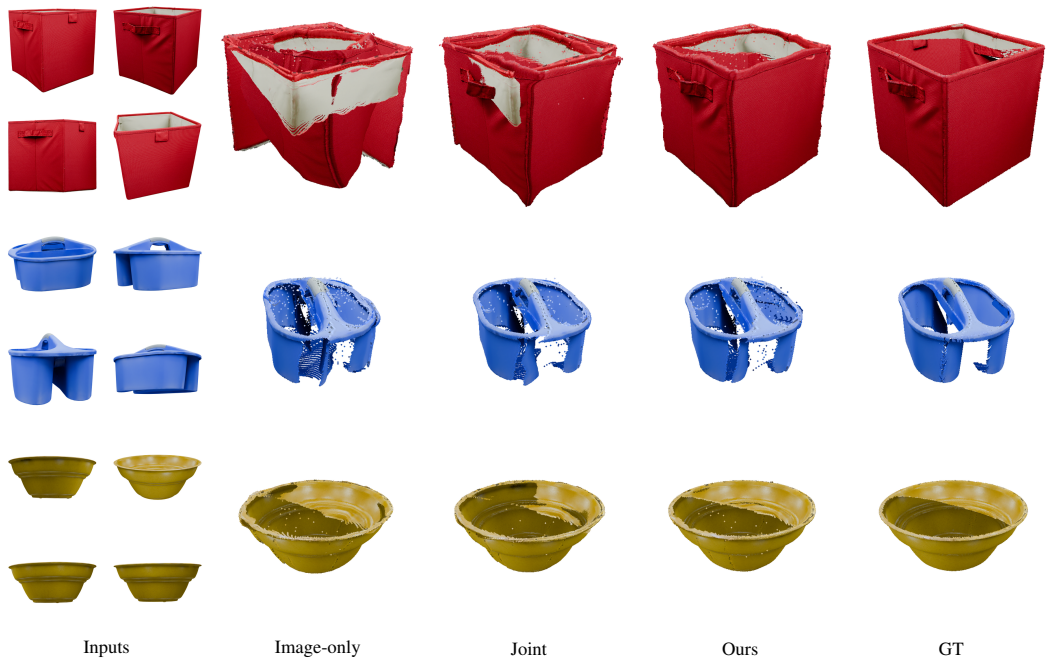

Figure 6: **Qualitative comparison of point-map reconstruction.** Each row shows one object. From left to right: four input views, point maps predicted by image-only VGGT, jointly trained VGGT, and structure-conditioned VGGT (ours), followed by ground-truth visible point maps.

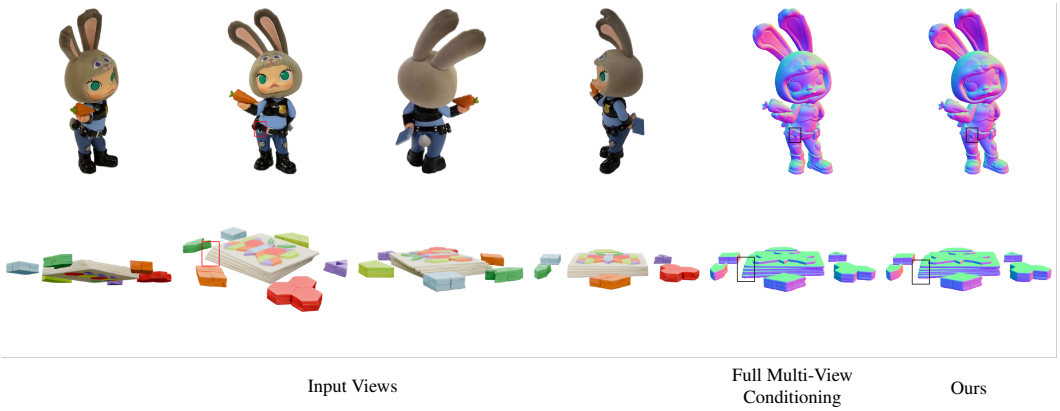

Figure 7: **Qualitative ablation of voxel-wise view selection and conditioning.** Each row shows four input views followed by rendered normal maps of shapes generated with full multi-view conditioning and our method, respectively. Both variants use the same predicted sparse structure. Red boxes highlight reference details in the input images, and black boxes mark the corresponding regions in the generated shapes.

A.2.4 Visualization of Voxel-Wise View Selection

Fig. 9 visualizes voxel-wise view selection from our model's actual inference results, including the generated sparse structures, predicted pixel-to-voxel correspondences, and selected views. Pixel support measures the amount of image evidence associated with a local 3D region and serves as a practical proxy for view informativeness. In these examples, the selected views expose the queried surface regions more directly and provide stronger pixel support than the alternatives. These results qualitatively support our use of pixel-to-voxel correspondences to identify informative views and provide region-specific visual evidence for detailed shape generation.

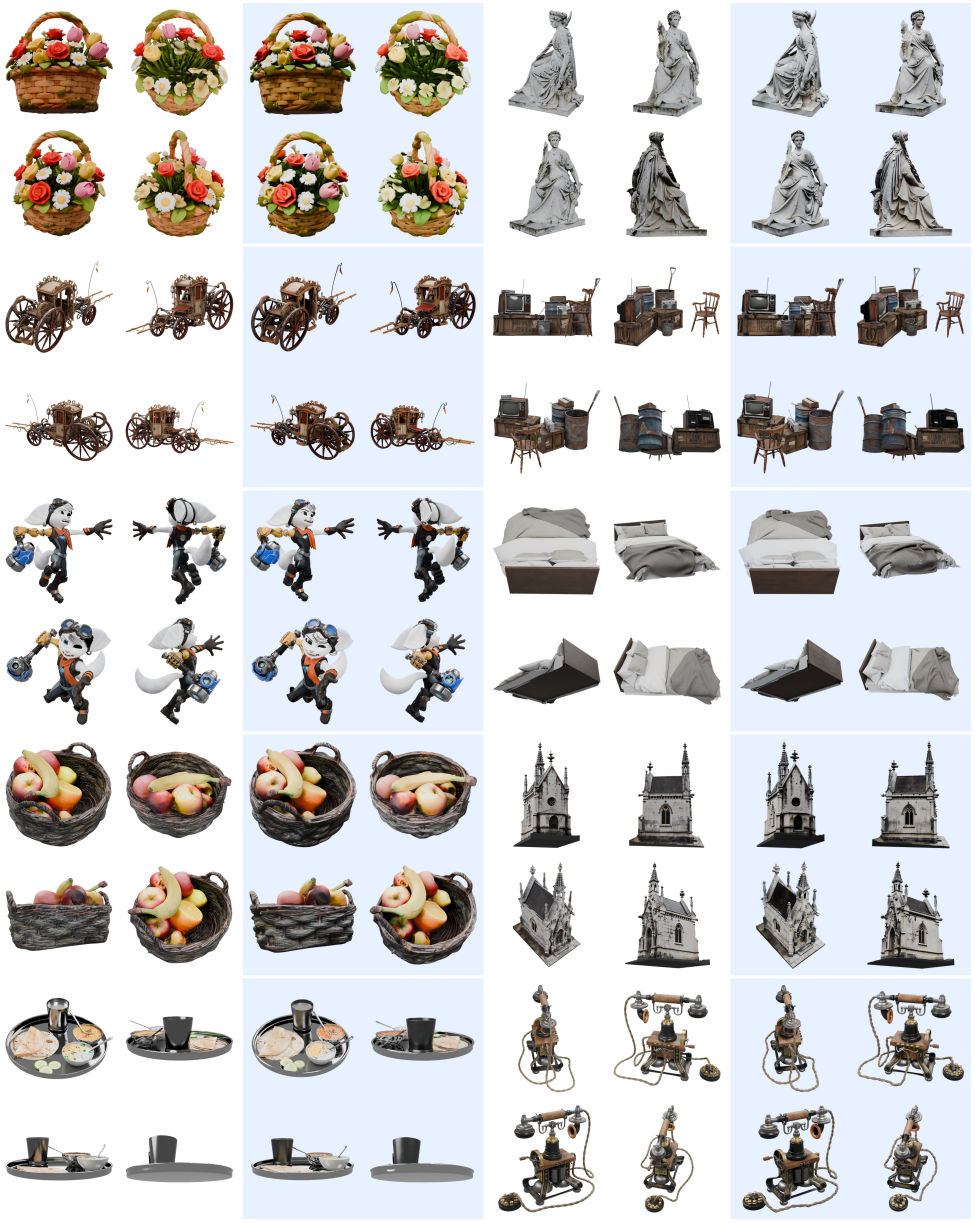

Figure 8: **Additional qualitative results on rendered 3D assets.** Each group shows four input views on a white background and the corresponding renderings of our generated mesh on a light-blue background. For direct visual comparison, we geometrically align the generated mesh with the ground-truth asset and render it using the same cameras as the input views.

A.2.5 Shared-Space Consistency between Reconstruction and Generation

Figures 10 and 11 visualize the reconstructed point clouds together with the generated meshes for *Structure-Conditioned VGGT* and *Jointly Trained VGGT*, respectively. All results are obtained directly from inference and shown in the shared model coordinate system, without object-specific alignment or registration. The close correspondence in the overlays demonstrates that the reconstructed point clouds and generated meshes reside in the same space and are directly spatially aligned.

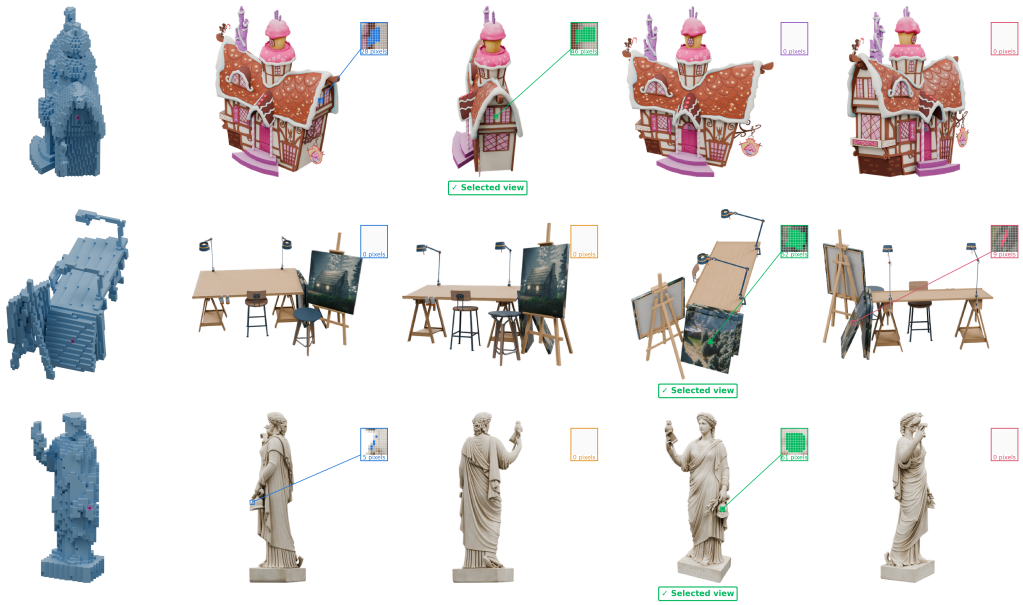

Figure 9: **Visualization of voxel-wise view selection from our inference results.** Each row shows a generated sparse structure with a query voxel highlighted in magenta, followed by four input views. Colored pixels and enlarged insets show the predicted pixel-to-voxel correspondences, with numbers indicating highlighted pixels within each inset. Green checkmarks identify the views selected by our model.



Figure 10: Shared-space consistency of *Structure-Conditioned VGGT*. Each row shows the reconstructed colored point cloud, the generated mesh, and their direct overlay from left to right. No object-specific alignment or registration is applied.



Figure 11: Shared-space consistency of *Jointly Trained VGGT*. Each row shows the reconstructed colored point cloud, the generated mesh, and their direct overlay from left to right. No object-specific alignment or registration is applied.