SUPRA-VOX: NATIVE VOXEL PROMPTS FOR SUPER-ALIGNED 3D ASSET GENERATION

Anonymous authors
Paper under double-blind review

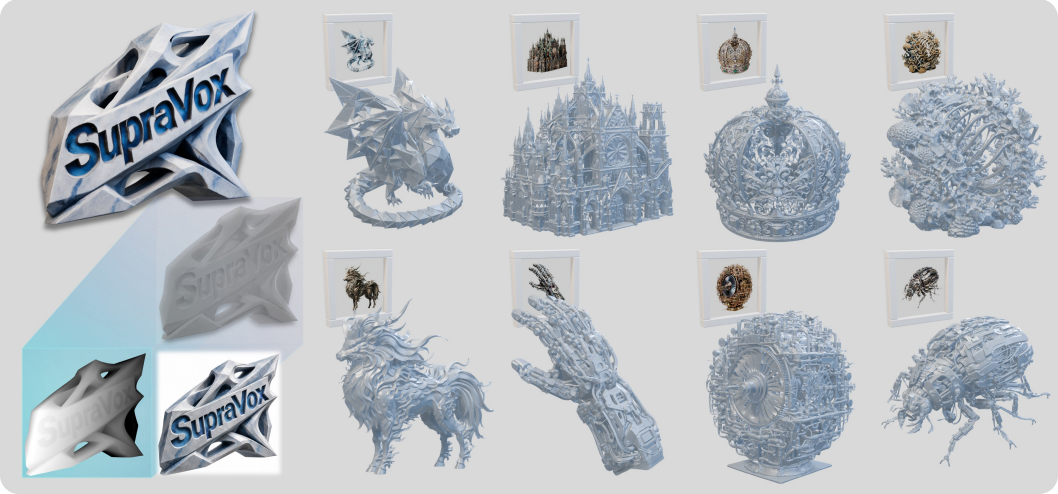


Figure 1: **SUPRA-Vox: toward super-aligned 3D generation that delivers *what you see is what you get*.** Assets generated by SUPRA-Vox.

ABSTRACT

Image-to-3D models can generate visually compelling assets, but often struggle to faithfully reproduce the geometry visible in the input view. In multi-stage 3D generation, missing parts or misplaced surfaces in the initial structure can persist in the final shape, even after later stages add fine details. We propose **SUPRA-Vox**, a surface-guided framework for sparse voxel generation toward *what you see is what you get* geometry. SUPRA-Vox encodes the visible surface reconstructed by a geometry foundation model into *Native Voxel Prompts* using the generator's pretrained variational autoencoder (VAE). These prompts remain available as separate conditions throughout denoising. Together with a visual hull mask, they guide the inpainting transformer as it jointly updates the complete voxel structure. The resulting structure is passed to a frozen downstream generator for detailed shape synthesis. We train SUPRA-Vox on a curated collection of high-quality 3D assets and evaluate it on two complex-geometry benchmarks we construct, Complex-Geo and DTC-Geo. Experiments on these benchmarks demonstrate state-of-the-art visible-surface reconstruction performance among the evaluated methods and improved overall geometric quality over native Pixal3D, demonstrating that better early structure benefits final geometry.

1 INTRODUCTION

Image-to-3D generation aims to create complete 3D assets from a reference image while faithfully reproducing the depicted object's shape and appearance. By drawing on 2D diffusion priors, score distillation sampling has advanced high-quality 3D content generation (Poole et al., 2023; Lin et al., 2023; Wang et al., 2023c). More recently, native 3D generators based on latent vector sets and sparse voxels have improved geometric quality and generation efficiency (Zhang et al., 2023; Zhao

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.

et al., 2023; Xiang et al., 2025; 2026), supporting product digitization, virtual environments, and embodied AI (Yang et al., 2024c; Wang et al., 2024c). However, visual plausibility does not ensure faithful geometry: a model may omit visible parts, alter an object's silhouette, or introduce local structures unsupported by the image (Li et al., 2026a; Xia et al., 2026). Our goal is *what you see is what you get* geometry: faithfully preserving the visible shape, part layout, and surface structure revealed by the reference image, while plausibly completing unobserved regions.

Achieving this goal requires accurate structure during the early stages of generation. In sparse-voxel frameworks, generation begins by predicting voxel locations, which then guide the synthesis of detailed geometry (Xiang et al., 2025; 2026). Recent methods improve input-view alignment through joint shape-and-pose modeling (Chen et al., 2026b; Huang et al., 2026) or projection-based image conditioning (Huang et al., 2026; Li et al., 2026a). These mechanisms help align objects with the camera and associate image evidence with 3D locations, but the generator must still infer where the visible surface lies along each viewing ray. Errors in this inference can lead to missing parts or misplaced surfaces. This motivates the use of explicit visible-surface observations to guide sparse voxel generation (Fig. 2).

Recent advances in monocular geometry estimation provide dense, pixel-aligned surface predictions from a single image (Wang et al., 2025c;d; Kong et al., 2026). Reconstruction priors have also been incorporated into 3D generation through geometry-aware feature conditioning (Chang et al., 2026), latent inpainting (Xia et al., 2026), and geometry-informed noise initialization (Liu et al., 2026). Building on these approaches, we explore how explicit surface estimates can complement pixel-aligned image conditioning in sparse voxel generation. Pixel-aligned image features associate visual evidence with 3D locations (Li et al., 2026a), while reconstructed surfaces provide explicit estimates of where the visible geometry lies. Bringing these signals into a shared spatial representation allows them to jointly guide the emerging structure. We introduce **Native Voxel Prompts** to represent reconstructed visible surfaces in the generator's own sparse-structure latent space. We align the surface with the generation space, the bounded 3D domain used by the generator, and encode its voxelized geometry with the pretrained VAE. The resulting prompts provide spatially aligned surface evidence that complements image features during structure generation.

With these prompts, we propose **SUPRA-Vox**, a framework for surface-guided sparse voxel generation. Our central idea is to keep surface observations available throughout generation while allowing the complete structure to evolve. As shown in Fig. 3, the SUPRA-Vox Inpainting Transformer combines Native Voxel Prompts with pixel-aligned image features, maintaining the prompts as a separate condition throughout denoising. A visual hull mask identifies candidate regions behind the visible surface, providing spatial cues for completion. Together, these conditions guide the joint generation of observed and unobserved regions. We find that explicit surface guidance at this early stage improves visible-surface fidelity and overall geometry in the final mesh.

To train and evaluate this framework on complex geometry, we curate a high-quality collection of 3D training assets and construct two evaluation sets, Complex-Geo and DTC-Geo. The latter is curated from the existing DTC dataset. On these benchmarks, SUPRA-Vox achieves state-of-the-art visible-surface reconstruction performance among the compared methods. Guided by depth-derived **Native Voxel Prompts**, it closely reproduces the geometry revealed by the reference image, advancing our goal of *what you see is what you get* 3D generation. Improving the sparse voxel structure also improves the final mesh produced by the unchanged downstream generator. We will release our code and both evaluation sets to support further research.

Our contributions are summarized as follows:

- We propose **SUPRA-Vox**, a surface-guided sparse voxel generation framework. Its Inpainting Transformer combines persistent surface conditions with pixel-aligned image features, guiding complete-structure generation while allowing observed and unobserved regions to evolve jointly.
- We introduce **Native Voxel Prompts** to complement pixel-aligned visual conditioning with explicit estimates of visible-surface locations. By expressing reconstructed surfaces in the generator's native structure space, these prompts allow image evidence and observed geometry to jointly guide the formation of the complete structure.
- We curate high-quality fine-tuning data and establish two challenging geometry benchmarks, Complex-Geo and DTC-Geo. SUPRA-Vox achieves state-of-the-art visible-surface

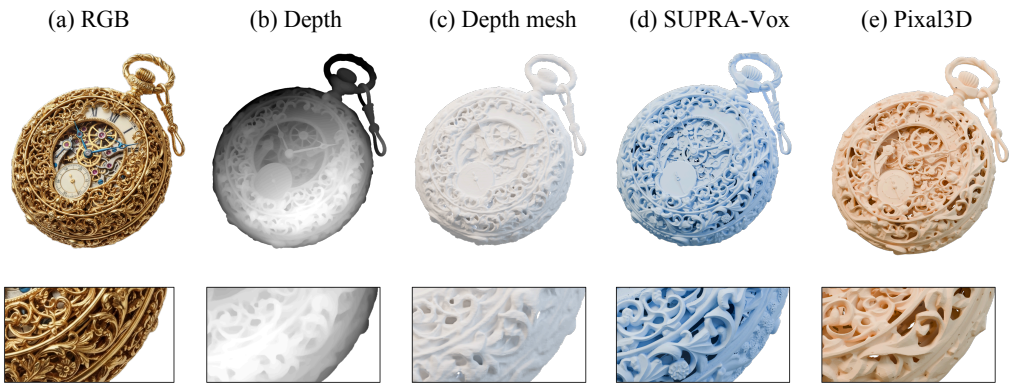


Figure 2: **Depth guidance improves 3D alignment with the input image.** (a) Reference image; (b) predicted depth; (c) depth-derived mesh; (d) SUPRA-Vox; (e) Pixal3D. Insets show corresponding surface details.

reconstruction performance among the evaluated methods on both benchmarks and improves aligned whole-mesh geometry over native Pixal3D on Complex-Geo.

2 Related Work

3D generative models. 3D generative methods mainly produce a complete 3D asset from a single image or a text prompt. Early work learned object-level distributions over implicit fields, points, voxels, or textured meshes (Chen & Zhang, 2019; Zhou et al., 2021; Nichol et al., 2022; Jun & Nichol, 2023; Gao et al., 2022), and thereby established samplable complete-shape priors. Score distillation then distilled large 2D diffusion priors into per-instance 3D optimization (Poole et al., 2023; Wang et al., 2023a; Metzer et al., 2023; Lin et al., 2023; Chen et al., 2023; Wang et al., 2023c; Tang et al., 2023; 2024b; Liang et al., 2024), substantially improving appearance realism and open-vocabulary generation. Building on this, multiview diffusion and feed-forward reconstruction (Liu et al., 2023; Shi et al., 2024; Liu et al., 2024; Shi et al., 2023; Long et al., 2024; Hong et al., 2024; Tochilkin et al., 2024; Wang et al., 2024d; Tang et al., 2024a; Xu et al., 2024; Wu et al., 2024c; Boss et al., 2025) strengthened cross-view consistency. In parallel, structured native 3D representations—vector sets, triplanes, hierarchical sparse latents, and surface-aware voxels— (Zhang et al., 2023; Zhao et al., 2023; Wang et al., 2023b; Lan et al., 2024; Wu et al., 2024d; Zhang et al., 2024; Ren et al., 2024; Xiong et al., 2025; Chen et al., 2025; He et al., 2025; Xiang et al., 2025; 2026; Wu et al., 2026; Wang et al., 2026; Zhao et al., 2026) enabled complete-asset generation to scale up inside the generator's own 3D representation, including direct synthesis of mesh vertices and topology. These advances strengthen the prior over complete 3D assets. SUPRA-Vox complements this prior with Native Voxel Prompts that provide explicit evidence of the input-visible geometry.

3D reconstruction models. Image-driven 3D reconstruction has moved from per-scene optimization to feed-forward geometry prediction. Classical multi-view stereo, neural radiance fields, implicit surfaces, and static Gaussians (Schonberger & Frahm, 2016; Yao et al., 2018; Mildenhall et al., 2020; Wang et al., 2021; Kerbl et al., 2023) recover cameras and scene structure from correspondence and photometric consistency. The same explicit-representation line was later extended to dynamic and high-dynamic-range scenes (Wu et al., 2024b; Zhang et al., 2025a; Wu et al., 2024a), enabling real-time novel-view synthesis and extractable time-varying geometry. Pixel-aligned geometric regression and unified visual geometry models (Wang et al., 2024b; Leroy et al., 2024; Wang et al., 2024a; Wang & Agapito, 2025; Yang et al., 2025; Wang et al., 2025b; Cabon et al., 2025; Wang et al., 2025a; Lin et al., 2026) then support camera and dense-structure recovery from uncalibrated or arbitrarily many views. Reconstruction-guided generative methods further combine multi-view reconstruction priors with learned 3D shape priors to improve consistency in global structure and local detail while completing sparsely observed objects (Chang et al., 2026). In parallel, monocular estimation progressed from relative depth to metric, camera-aware, generative, and affine-invariant pointmap prediction (Ranftl et al., 2021; 2022; Bhat et al., 2023; Yin et al., 2023; Hu et al., 2024;

Piccinelli et al., 2024; Yang et al., 2024a;b; Bochkovskii et al., 2025; Ke et al., 2024; Fu et al., 2024; Wang et al., 2025c; Kong et al., 2026), providing useful estimates of visible surface layout from a single image. These estimates primarily capture overall geometry; thin structures, sharp boundaries, and high-frequency surface detail can remain incomplete. We encode their predictions as Native Voxel Prompts to guide coarse sparse structure, with finer geometry synthesized by the downstream generative model.

Prompting 3D generative models with observed priors. Partial-geometry-conditioned 3D generation first developed along point-cloud completion (Yuan et al., 2018; Tchapmi et al., 2019; Huang et al., 2020; Wen et al., 2020; Liu et al., 2020; Wang et al., 2020; Xie et al., 2020), showing that a local observation can constrain a complete object. Probabilistic modeling, sparse discrete representations, and conditional generative models (Zhou et al., 2021; Pan et al., 2021; Yan et al., 2022; Cheng et al., 2023; Chu et al., 2023) then used a generative complete-shape prior for high-quality point-cloud completion. Recent work instead reuses large pretrained 2D or 3D generative priors (Yu et al., 2023; Li et al., 2025a; Yan et al., 2026; Xia et al., 2026): training-free spatial replacement and latent correction, or learned mask-aware native-state inpainting, allow a pretrained complete-shape prior to keep working under a partial observation. For multi-view inputs, reconstruction priors can also guide diffusion initialization through geometry-based noise inversion and confidence-aware spatial noise modulation, balancing fidelity to observed surfaces with completion of unobserved regions (Liu et al., 2026). At scene scale, structured 3D latent block inpainting has similarly been used to extend existing geometry with context-aware generated regions (Li et al., 2026b). Meanwhile, semantic and pixel-aligned image conditions (Chen et al., 2026a; Li et al., 2026a) further strengthen the correspondence between the input view and 3D space. Points-to-3D (Xia et al., 2026) encodes partial geometry into native structure latents and combines them with noise for mask-aware inpainting and boundary refinement. SUPRA-Vox supplies the observation as a separate condition throughout denoising and uses a view-derived mask to identify space requiring completion.

3 Preliminaries

Pixal3D (Li et al., 2026a) generates assets in an input-view-aligned frame using projected image features $\mathbf{h}_{\mathcal{I}}$. We focus on its sparse voxel generation, which determines voxel locations for subsequent shape and texture synthesis. Following TRELLIS.2 (Xiang et al., 2026), a sparse-structure VAE encodes a binary 64^3 surface-occupancy volume $\mathcal{V}_{\mathcal{C}}$, where $\mathcal{C}$ is the set of surface-voxel coordinates:

$$\mathbf{z}_{\mathrm{SS}} = \mathcal{E}_{\mathrm{SS}}(\mathcal{V}_{\mathcal{C}}), \qquad \widetilde{\mathcal{C}} = \mathrm{supp}[\mathcal{D}_{\mathrm{SS}}(\mathbf{z}_{\mathrm{SS}}) > 0]\,, \tag{1}$$

where $\mathbf{z}_{\mathrm{SS}} \in \mathbb{R}^{8\times16\times16\times16}$ and $\widetilde{\mathcal{C}}$ is the reconstructed voxel coordinate set obtained by thresholding decoded occupancy logits at zero.

An image-conditioned Flow Transformer predicts a velocity field over the noisy latent $\mathbf{x}_t$:

$$\mathbf{v}_\theta^{\mathrm{base}}(\mathbf{x}_t, t, \mathbf{h}_{\mathcal{I}}) = \mathcal{F}_\theta^{\mathrm{base}}(\mathbf{x}_t, t, \mathbf{h}_{\mathcal{I}}). \tag{2}$$

The training path is $\mathbf{x}_t = (1-t)\mathbf{z}_{\mathrm{SS}} + [\sigma_{\min} + (1-\sigma_{\min})t]\boldsymbol{\epsilon}$, where $t \in [0,1]$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, and $\sigma_{\min}$ is the noise floor. Sampling integrates this velocity from $t = 1$ to $t = 0$ to obtain $\widehat{\mathbf{z}}_{\mathrm{SS}} = \mathcal{G}_{\mathrm{SS}}^{\mathrm{base}}(\boldsymbol{\epsilon}, \mathbf{h}_{\mathcal{I}})$. Thresholding the decoded occupancy yields coordinates $\widehat{\mathcal{C}}$. The frozen downstream pipeline produces the asset $\mathcal{A} = \Phi(\widehat{\mathcal{C}})$, with image and noise inputs omitted for brevity.

4 Method

4.1 SUPRA-Vox Framework

Given the input image $\mathcal{I}$, SUPRA-Vox first prepares **Native Voxel Prompts** from the predicted visible surface, then uses the SUPRA-Vox Inpainting Transformer to generate a complete sparse structure, as shown in Fig. 3:

$$\begin{aligned}
&\text{(i) prepare:} && (\mathbf{z}_{\mathrm{obs}}^{\mathrm{str}}, \mathbf{m}_{\mathrm{obs}}) = \mathcal{S}\big(\mathcal{P}_{\mathrm{geo}}(\mathcal{I}); \mathbf{K}, \mathbf{T}, \mathcal{G}\big),\\
&\text{(ii) inpaint:} && \widehat{\mathbf{z}}_{\mathrm{SS}} = \mathcal{G}_{\mathrm{SS}}^{\mathrm{SV}}(\boldsymbol{\epsilon}, \mathbf{h}_{\mathcal{I}}, \mathbf{z}_{\mathrm{obs}}^{\mathrm{str}}, \mathbf{m}_{\mathrm{obs}}),\\
&\text{(iii) read out:} && \mathcal{A} = \Phi(\widehat{\mathcal{C}}), \qquad \widehat{\mathcal{C}} = \mathrm{supp}[\mathcal{D}_{\mathrm{SS}}(\widehat{\mathbf{z}}_{\mathrm{SS}}) > 0]\,.
\end{aligned} \tag{3}$$

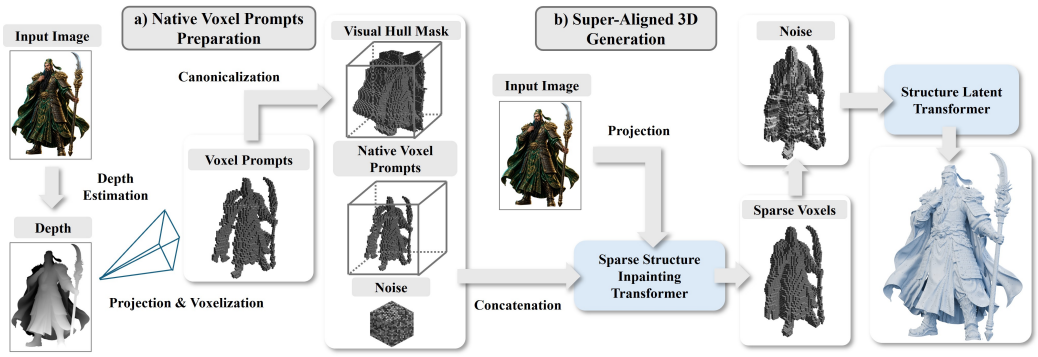


Figure 3: **Overview of SUPRA-Vox.** Native Voxel Prompts guide sparse structure generation, followed by detailed shape synthesis with the frozen downstream generator.

Step (i) corresponds to Native Voxel Prompts Preparation in Fig. 3(a). The operator $\mathcal{S}$ aligns and encodes the visible surface predicted by $\mathcal{P}_{\text{geo}}$ into the Native Voxel Prompt $\mathbf{z}^{\text{str}}_{\text{obs}}$ and constructs the visual hull mask $\mathbf{m}_{\text{obs}}$ (Section 4.2). The camera intrinsics $\mathbf{K}$, pose $\mathbf{T}$, and canonicalization transform $\mathcal{G}$ maintain correspondence between the input view and the generated structure. In step (ii), $\mathcal{G}^{\text{SV}}_{\text{SS}}$ samples with the SUPRA-Vox Inpainting Transformer conditioned on the prompts, mask, and image features (Section 4.3). Step (iii) decodes the generated voxel structure and applies the frozen downstream pipeline Φ; together, these steps form the generation branch in Fig. 3(b).

4.2 Native Voxel Prompts Preparation

The preparation operator $\mathcal{S}$ converts the reconstructed visible surface into a Native Voxel Prompt $\mathbf{z}^{\text{str}}_{\text{obs}}$ and an accompanying visual hull mask $\mathbf{m}_{\text{obs}}$.

$$\mathcal{I} \xrightarrow{\mathcal{P}_{\text{geo}}} \mathbf{P} \xrightarrow{\mathbf{T},\mathcal{G}} \widetilde{\mathbf{P}} \xrightarrow{\text{voxelize}} \mathcal{V}_{\text{obs}} \xrightarrow{\mathcal{E}_{\text{SS}}} \mathbf{z}^{\text{str}}_{\text{obs}}, \tag{4}$$

with the mask $\mathbf{m}_{\text{obs}}$ branching off the canonical points $\widetilde{\mathbf{P}}$.

Native Voxel Prompts. We use a pretrained monocular geometry model $\mathcal{P}_{\text{geo}}$—MoGe-3 (Kong et al., 2026) by default—to obtain a pixel-aligned visible-surface point map $\mathbf{P}$ over the valid foreground pixels $\Omega_{\mathcal{I}}$. Its camera-space points are mapped into the generator's coordinate frame:

$$\mathbf{P} = \mathcal{P}_{\text{geo}}(\mathcal{I}), \qquad \widetilde{\mathbf{P}} = \mathcal{G}(\mathbf{T}\mathbf{P}). \tag{5}$$

For estimators that output only camera-Z depth $\mathbf{D}$, the same interface is obtained by $\mathbf{P}(\mathbf{u}) = \mathbf{D}(\mathbf{u})\mathbf{K}^{-1}\bar{\mathbf{u}}$, where $\bar{\mathbf{u}}$ is the homogeneous pixel coordinate. We calibrate the canonical bounding box using the geometric prior predicted by native Pixal3D, defining $\mathcal{G} = \mathcal{G}_{\text{prior}}$; details are provided in Appendix B.3.

We express the aligned surface $\widetilde{\mathbf{P}}$ on the generator's 64^3 voxel grid. The operator $\mathcal{R}_{64}$ quantizes points to cells, merges duplicates, and discards disconnected outliers. Then the surviving cells are the observed coordinates $\mathcal{C}_{\text{obs}}$. The frozen encoder $\mathcal{E}_{\text{SS}}$ of (1) then lifts them into latent space, using the binary occupancy volume $\mathcal{V}_{\text{obs}}$ induced by these coordinates:

$$\mathcal{C}_{\text{obs}} = \mathcal{R}_{64}(\widetilde{\mathbf{P}}), \qquad \mathbf{z}^{\text{str}}_{\text{obs}} = \mathcal{E}_{\text{SS}}(\mathcal{V}_{\text{obs}}). \tag{6}$$

Reusing the native VAE encoder places the observed surface in the same latent space as the complete structure to be generated.

Visual Hull Mask. The mask marks candidate completion regions behind the visible surface within the foreground projection. Rasterizing $\widetilde{\mathbf{P}}$ onto the input view gives a first-hit depth D_{hit}. A 64^3 cell $\mathbf{x}$ is selected when it projects inside the foreground at $\pi(\mathbf{x})$ and its camera-space depth $z(\mathbf{x})$ falls strictly behind that first hit:

$$m_{64}(\mathbf{x}) = \mathbf{1}\big[\pi(\mathbf{x}) \in \Omega_{\mathcal{I}} \ \wedge \ z(\mathbf{x}) > D_{\text{hit}}(\pi(\mathbf{x}))\big], \qquad \mathbf{m}_{\text{obs}} = \text{NN}_{16}(m_{64}). \tag{7}$$

where $\mathbf{1}[\cdot]$ is the indicator and NN_{16} is nearest-neighbor downsampling from 64^3 onto the 16^3 Flow grid. The mask identifies candidate completion regions and serves as a spatial condition for the transformer.

The Native Voxel Prompt $\mathbf{z}_{\mathrm{obs}}^{\mathrm{str}}$ provides surface content, while the mask $\mathbf{m}_{\mathrm{obs}}$ supplies complementary spatial information. Both depend on the canonical box and the camera used in the lifting. Fig. 5 shows how the box fit changes the generated structure.

4.3 SUPRA-Vox Inpainting Transformer

The SUPRA-Vox Inpainting Transformer conditions sparse structure generation on Native Voxel Prompts and the visual hull mask while retaining Pixal3D's original image-feature injection mechanism. At each denoising step, the prompt $\mathbf{z}_{\mathrm{obs}}^{\mathrm{str}}$ and mask $\mathbf{m}_{\mathrm{obs}}$ are concatenated with the noisy structure state $\mathbf{x}_t$ along the channel dimension and passed through an expanded input projection. The transformer $\mathcal{F}_\theta$ predicts the velocity conditioned on these inputs, time t, and image features $\mathbf{h}_{\mathcal{I}}$. The prompt and mask remain fixed throughout sampling, while the predicted velocity updates the complete structure state. This allows visible-surface evidence to guide both geometric alignment and completion of unobserved regions. We initialize from pretrained Pixal3D weights and zero-initialize the additional input-channel weights. Training follows the original flow-matching objective:

$$\mathcal{L}_{\mathrm{SV}} = \mathbb{E}\left[\left\|\mathcal{F}_\theta([\mathbf{x}_t; \mathbf{m}_{\mathrm{obs}}; \mathbf{z}_{\mathrm{obs}}^{\mathrm{str}}], t, \mathbf{h}_{\mathcal{I}}) - \mathbf{v}^\star\right\|_2^2\right], \tag{8}$$

where $\mathbf{v}^\star = (1 - \sigma_{\min})\boldsymbol{\epsilon} - \mathbf{z}_{\mathrm{SS}}$ is the target velocity for the training path defined in Section 3.

5 Experiments

5.1 Experimental Setup

Datasets. We fine-tune on our HQ100k training set of 103,549 objects and 205,770 views, and evaluate on Complex-Geo (115 assets) and DTC-Geo (100 assets from Digital Twin Catalog, DTC (Dong et al., 2025)), both targeting complex geometry. Data sources, selection criteria, and benchmark construction are detailed in Appendices B.1 and B.4.

Implementation details. We fine-tune Pixal3D's sparse voxel generator with the VAE and downstream models frozen. Prompts use mesh-rendered depth during training and MoGe-3 (ViT-L) at inference. SUPRA-Vox uses MoGe-3-estimated camera intrinsics; baselines follow their respective inference frameworks. Sampling uses 12 Euler steps without classifier-free guidance; details are in Appendix A.

5.2 Main Results

We compare with LaS-Comp (Yan et al., 2026), Step1X-3D (Li et al., 2025b), Direct3D-S2 (Wu et al., 2025), SAM 3D (Chen et al., 2026b), UniLat3D (Wu et al., 2026), Hunyuan3D 2.1 (Team Hunyuan3D et al., 2025), TRELLIS.2 (Xiang et al., 2026), CUPID (Huang et al., 2026), and native Pixal3D (Li et al., 2026a). We assess visible surfaces using normal-map PSNR, SSIM, and LPIPS (Zhang et al., 2018), angular error (including boundaries), angular accuracy, and silhouette IoU; whole-mesh metrics are Chamfer L_1, F-score, Earth Mover's Distance (EMD), and vIoU. Uni3D (Zhou et al., 2024) and ULIP-2 (Xue et al., 2024) measure image–3D agreement. SUPRA-Vox with GT-depth prompts serves as an oracle observation control; evaluation details are provided in Appendix B.5.

SUPRA-Vox achieves the best normal PSNR, SSIM, and angular accuracy among the compared non-oracle methods on both datasets (Table 1). These consistent gains show that Native Voxel Prompts improve agreement with reference-visible geometry. On DTC-Geo, native Pixal3D retains the best silhouette IoU and LPIPS among non-oracle methods.

On Complex-Geo, SUPRA-Vox improves all reported aligned whole-mesh geometry metrics and image–3D agreement scores over native Pixal3D (Table 2). This shows that early structural guidance benefits the final geometry with the downstream generator frozen. Qualitative comparisons in Fig. 4 complement these results; additional examples appear in Appendix C.

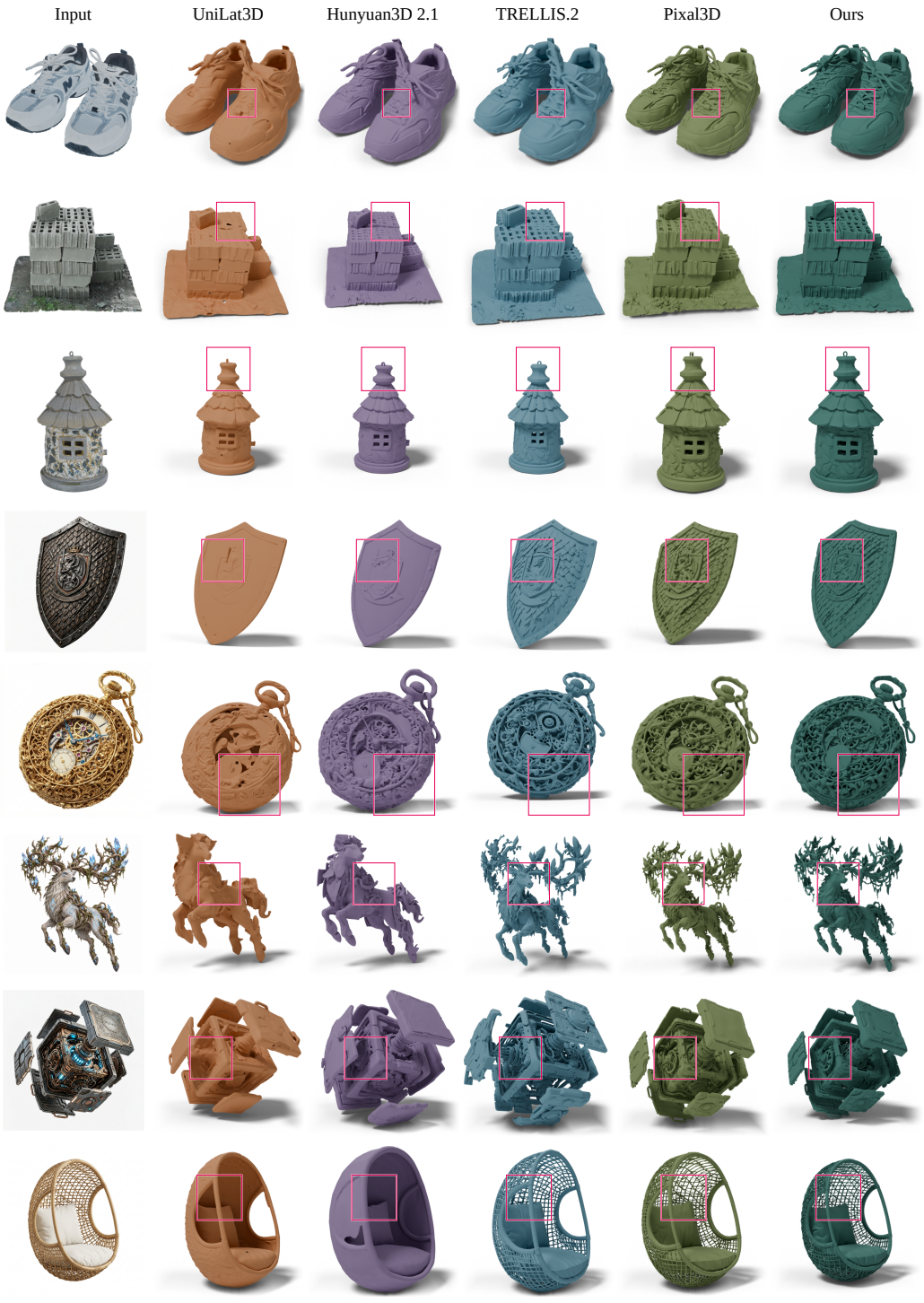


Figure 4: **Qualitative comparison** on evaluation-set and in-the-wild images. Ours denotes SUPRA-Vox; red boxes highlight local geometry.

The qualitative comparisons highlight thin structures, openwork surfaces, and repeated parts. Native Voxel Prompts anchor sparse support to visible geometry, while the frozen downstream model supplies local detail. These examples complement the quantitative evaluation of reference-view fidelity and complete-shape geometry.

Table 1: **SUPRA-Vox achieves the best normal PSNR, SSIM, and angular accuracy among the evaluated non-oracle methods on both benchmarks.**

Method	Silhouette	Normal-map similarity			Angular error (°)			Angular accuracy (%)		
	IoU ↑	PSNR ↑	SSIM ↑	LPIPS ↓	Mean ↓	Median ↓	Boundary ↓	11.25° ↑	22.5° ↑	30° ↑
Complex-Geo										
LaS-Comp	0.534	18.37	0.282	0.227	51.20	48.26	54.18	5.7	16.6	26.0
Step1X-3D	0.788	20.92	0.838	0.133	38.12	31.84	42.12	18.4	37.6	48.2
Direct3D-S2	0.805	20.81	0.835	0.129	39.42	33.78	43.09	16.7	34.7	45.7
SAM 3D	0.802	20.91	0.838	0.134	37.84	31.23	41.43	20.1	38.6	48.9
UniLat3D	0.842	21.77	0.850	0.111	33.62	26.65	37.39	24.6	44.9	55.4
Hunyuan3D 2.1	0.841	21.69	0.849	0.108	34.14	27.27	37.62	23.7	43.8	54.6
TRELLIS.2	0.845	21.66	0.847	0.103	34.48	27.65	37.95	24.0	43.6	53.8
CUPID	0.894	22.23	0.858	0.095	32.13	24.91	35.08	25.6	47.2	58.1
Pixal3D	0.945	24.11	0.892	0.058	26.04	19.62	28.28	31.3	57.0	68.4
SUPRA-Vox	**0.956**	24.76	0.900	0.056	24.02	17.94	26.48	34.3	61.0	72.1
SUPRA-Vox (GT-depth)	0.955	**25.03**	**0.903**	**0.053**	**22.61**	**16.02**	**25.11**	**40.5**	**64.3**	**74.0**
DTC-Geo										
LaS-Comp	0.589	18.78	0.324	0.204	49.69	46.41	53.48	5.7	16.6	26.3
Step1X-3D	0.676	19.92	0.837	0.164	43.10	37.44	47.15	13.0	29.0	39.7
Direct3D-S2	0.868	21.59	0.853	0.109	34.21	26.66	36.21	22.8	42.1	53.4
SAM 3D	0.674	19.91	0.840	0.163	43.59	37.55	47.64	14.1	29.3	39.4
UniLat3D	0.694	20.02	0.837	0.160	41.54	35.09	46.01	15.7	32.0	42.3
Hunyuan3D 2.1	0.680	20.03	0.839	0.158	42.43	36.69	47.12	14.3	30.6	41.1
TRELLIS.2	0.680	19.79	0.832	0.164	42.91	37.13	47.33	14.8	30.2	40.3
CUPID	0.933	23.37	0.880	0.078	27.29	20.34	27.78	30.0	53.6	65.3
Pixal3D	**0.973**	25.58	0.910	0.051	21.65	16.66	21.89	34.4	64.4	76.6
SUPRA-Vox	0.968	26.17	0.917	0.053	19.54	15.03	20.39	38.1	69.2	80.9
SUPRA-Vox (GT-depth)	0.971	**26.75**	**0.920**	**0.050**	**17.53**	**12.59**	**18.70**	**46.4**	**74.0**	**83.7**

Table 2: **SUPRA-Vox improves whole-mesh geometry and image–3D agreement over native Pixal3D on Complex-Geo.**

Method	Geometry				Image–3D	
	CD (L_1) ↓	F-score (%) ↑	EMD (10^{-2}) ↓	vIoU ↑	Uni3D ↑	ULIP-2 ↑
SAM 3D	0.0628	16.63	11.16	0.088	34.13	37.64
Pixal3D	0.0411	24.63	8.68	0.134	32.06	37.02
SUPRA-Vox	0.0386	25.60	8.64	0.137	33.98	38.75
SUPRA-Vox (GT-depth)	**0.0269**	**41.19**	**7.69**	**0.225**	**34.87**	**40.15**

In-the-wild and human evaluation. SUPRA-Vox leads both automatic metrics on 50 in-the-wild images (Table 3). Ten experienced evaluators compare anonymous videos for 15 cases with randomized left–right placement. Align./Struct. report our win rates against each baseline, counting ties in the denominator (protocol: Appendix B.6).

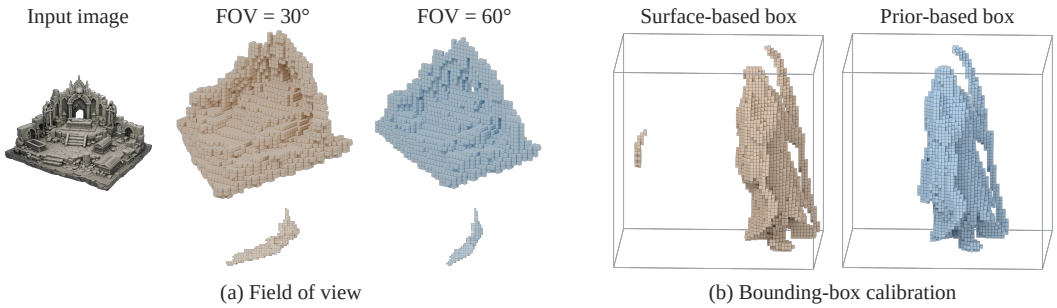
Table 3: **In-the-wild agreement and human preferences.**

Method	Uni3D ↑	ULIP-2 ↑	Our wins (%) Align. ↑	Our wins (%) Struct. ↑
UniLat3D	25.96	11.04	100.0	94.6
Hunyuan3D 2.1	34.71	12.64	86.8	84.2
TRELLIS.2	36.71	12.41	71.1	63.2
Pixal3D	36.45	12.77	54.1	51.4
SUPRA-Vox	**36.81**	**12.85**	–	–

5.3 Ablation Studies

Table 4 varies the conditions $(\mathbf{z}_{\text{obs}}^{\text{str}}, \mathbf{m}_{\text{obs}})$, geometry estimator $\mathcal{P}_{\text{geo}}$, and canonicalization $\mathcal{G}$ on Complex-Geo Subset with weights fixed. The default uses MoGe-3, a visual hull mask, and prior-based calibration. All variants share the downstream generator and sampling settings. Meshes are aligned by a similarity transform, with F-score evaluated at 2% of the ground-truth bounding-box diagonal.

Early structure shapes final geometry. We replace predicted sparse structure with complete ground-truth occupancy while keeping the downstream generator fixed: $\mathcal{A}_{\text{oracle}} = \Phi(\mathcal{C}_{\text{gt64}})$. The resulting improvement in whole-shape accuracy demonstrates the importance of establishing accurate structure before detail synthesis.



(a) Field of view (b) Bounding-box calibration

Figure 5: **Effect of FOV and bounding-box calibration.** (a) Fixed image and depth, varying lifting FOV. (b) Fixed observed voxels, varying calibration.

Explicit surface conditions improve fidelity. We set $(\mathbf{z}^{\text{str}}_{\text{obs}}, \mathbf{m}_{\text{obs}}) \leftarrow (\mathbf{0}, \mathbf{0})$ throughout denoising, with all other inputs and settings fixed. All four scores worsen, supporting explicit surface conditioning (Appendix B.2).

Observation quality affects surface fidelity. Replacing MoGe-3 with MoGe-2 (Wang et al., 2025d) rebuilds prompts and masks under the same rules. Accuracy declines modestly, while GT-depth prompts improve surface fidelity (Table 1), indicating sensitivity to observation quality.

Table 4: **Observation quality and spatial calibration affect final geometry.**

Setting	PSNR↑	SSIM↑	F@2%↑	Chamfer↓
SUPRA-Vox	**17.05**	**0.960**	**0.786**	**0.0158**
$(\mathbf{z}^{\text{str}}_{\text{obs}}, \mathbf{m}_{\text{obs}}) \leftarrow (\mathbf{0}, \mathbf{0})$	15.48	0.942	0.757	0.0167
$\mathbf{m}_{\text{obs}} \leftarrow \mathbf{m}_{\text{surf}}$	17.01	0.959	0.782	0.0161
$\mathcal{P}_{\text{geo}} \leftarrow \mathcal{P}_{\text{MoGe-2}}$	16.76	0.957	0.768	0.0166
$\mathcal{G} \leftarrow \mathcal{G}_{\text{surface}}$	17.02	0.960	0.654	0.0241
$\widehat{\mathcal{C}} \leftarrow \mathcal{C}_{\text{gt64}}$ (oracle)	17.93	0.966	0.975	0.0071

Camera lifting and spatial calibration. With RGB and depth fixed, changing the lifting FOV changes the reconstructed surface (Fig. 5a). Replacing prior-based calibration $\mathcal{G}_{\text{prior}}$ with a visible-surface box $\mathcal{G}_{\text{surface}}$ preserves reference-view fidelity but reduces whole-shape accuracy. Both variants rebuild prompts and masks with frozen weights. Prior-based calibration yields more coherent voxel structure (Fig. 5b).

Completion-region cues complement surface content. We replace the default mask $\mathbf{m}_{\text{obs}}$ from (7) with a surface-occupancy mask $\mathbf{m}_{\text{surf}} = \text{NN}_{16}(\mathcal{V}_{\text{obs}})$, keeping the surface prompt, camera, and generator fixed. The visual hull mask performs better, suggesting that completion-region cues complement observed surface content.

6 Discussion

Observation fidelity. Monocular prediction and voxelization can miss thin parts and fine surface variations. Incorrect Native Voxel Prompts can propagate to the final mesh (Appendix D). These observations provide an estimate of coarse visible structure rather than a complete description of surface detail. If they omit a part or misplace a surface, later detail synthesis need not correct the resulting structural error.

Fine geometry and additional views. Extending Native Voxel Prompts to fine geometry generation could improve local detail, which currently lacks direct high-resolution surface guidance. Complementary views could resolve unseen geometry and extend super-alignment to multi-view inputs (Liu et al., 2026).

7 Conclusion

We presented SUPRA-Vox, which encodes reconstructed surfaces as Native Voxel Prompts to guide sparse voxel generation. Together with a visual hull mask, these prompts condition the Inpainting Transformer throughout denoising. Experiments demonstrate improved visible-surface fidelity and whole-mesh geometry with a frozen downstream generator.

AI USE STATEMENT

Generative AI tools were used to assist with literature search and review, code implementation, experimental analysis, figure preparation, manuscript drafting and editing, and checking submission requirements against the conference guidelines. Generative AI was not used for research ideation. The data-curation pipeline also uses vision–language model annotations for asset categorization and scoring, as described in Appendix B.1. The authors are responsible for the final text, claims, and artifacts, including content prepared with AI assistance.

ETHICS STATEMENT

Single-image 3D generation can support creative workflows, but can also be misused to produce misleading assets or reproduce protected designs. Use and redistribution of input images, source assets, pretrained models, and generated content should respect applicable licenses, intellectual property rights, and privacy. Dataset selection and pretrained models may introduce biases in object coverage and generation quality; VLM-assisted curation can inherit annotation biases. Errors in unobserved regions and mesh connectivity limit its suitability for safety-critical applications.

REPRODUCIBILITY STATEMENT

Section 4 describes the model and conditioning formulation, and Section 5.1 summarizes the experimental setup. Appendix A specifies training and inference settings, including the optimizer, sampler, and random seed. Appendices B.1, B.3, and B.4 describe data selection, geometric calibration, and evaluation-set construction; Appendix B.5 details rendering, alignment, and metric computation. Ground-truth-assisted controls are identified separately from predicted-input evaluations.

REFERENCES

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

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

Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. SF3D: Stable fast 3D mesh reconstruction with UV-unwrapping and illumination disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16240–16250, 2025.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1050–1060, 2025.

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*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9278abf072b58caf21d48dd670b4c721-Abstract-Conference.html.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22189–22199, 2023.

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 *Computer Vision – ECCV 2026*, pp. 555–573. Springer, 2026a. doi: 10.1007/978-3-032-37490-5_31.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jia-Wei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3D: 3Dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_SAM_3D_3Dfy_Anything_in_Images_CVPR_2026_paper.html.

Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, Liang Pan, Dahua Lin, and Ziwei Liu. 3DTopia-XL: Scaling high-quality 3D asset generation via primitive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26576–26586, 2025.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5932–5941, 2019.

Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander Schwing, and Liang-Yan Gui. SDFusion: Multimodal 3D shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4456–4465, 2023.

Ruihang Chu, Enze Xie, Shentong Mo, Zhenguo Li, Matthias Nießner, Chi-Wing Fu, and Jiaya Jia. DiffComplete: Diffusion-based generative 3D shape completion. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75951–75966, 2023.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset and benchmarks for real-world 3D object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21126–21136, 2022.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35799–35813, 2023.

Zhao Dong, Ka Chen, Zhaoyang Lv, Hong-Xing Yu, Yunzhi Zhang, Cheng Zhang, Yufeng Zhu, Stephen Tian, Zhengqin Li, Geordie Moffatt, Sean Christofferson, James Fort, Xiaqing Pan, Mingfei Yan, Jiajun Wu, Carl Yuheng Ren, and Richard Newcombe. Digital twin catalog: A large-scale photorealistic 3D object digital twin dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 753–763, 2025.

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

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. GET3D: A generative model of high quality 3D textured shapes learned from images. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31841–31854, 2022.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14822–14833, 2025.

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

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Kaixuan Wang, Hao Chen, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3Dv2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 10579–10596, 2024.

Binbin Huang, Haobin Duan, Yiqun Zhao, Zibo Zhao, Yi Ma, and Shenghua Gao. CUPID: Generative 3D reconstruction via joint object and pose modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12741–12752, 2026. URL https://cupid3d.github.io/.

Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, and Xinyi Le. PF-Net: Point fractal network for 3D point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7659–7667, 2020.

Heewoo Jun and Alex Nichol. Shap-E: Generating conditional 3D implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. doi: 10.1145/3592433.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. 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*, pp. 16384–16393, 2024.

Lingyu Kong, Ruicheng Li, Ruicheng Wang, Sicheng Xu, Chengtang Yao, Jianfeng Xiang, and Jiaolong Yang. MoGe-3: Fine-detail monocular geometry estimation with self-guided sparse volumetric refinement. *arXiv preprint arXiv:2607.17967*, 2026.

Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. LN3Diff: Scalable latent neural fields diffusion for speedy 3D generation. In *European Conference on Computer Vision*, pp. 112–130, 2024.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, pp. 71–91, 2024.

An Li, Zhe Zhu, and Mingqiang Wei. GenPC: Zero-shot point cloud completion via 3D generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1308–1318, 2025a.

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 *ACM SIGGRAPH 2026 Conference Papers*, pp. 1–12. Association for Computing Machinery, 2026a. doi: 10.1145/3799902.3811175.

Sikuang Li, Chen Yang, Jiemin Fang, Taoran Yi, Jia Lu, Jiazhong Cen, Lingxi Xie, Wei Shen, and Qi Tian. WorldGrow: Generating infinite 3D world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6433–6441, 2026b. doi: 10.1609/aaai.v40i8.37571.

Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, Xiao Chen, Feipeng Tian, Jianxiong Pan, Zeming Li, Gang Yu, Xiangyu Zhang, Daxin Jiang, and Ping Tan. Step1X-3D: Towards high-fidelity and controllable generation of textured 3D assets. *arXiv preprint arXiv:2505.07747*, 2025b. URL https://arxiv.org/abs/2505.07747.

Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. LucidDreamer: Towards high-fidelity text-to-3D generation via interval score matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6517–6526, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 300–309, 2023.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4cd50120b6d7e8daff1749d6bbaa889-Abstract-Conference.html.

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. URL https://arxiv.org/abs/2609.11129.

Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11596–11603, 2020.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9264–9275, 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*, 2024.

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

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-NeRF for shape-guided generation of 3D shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12663–12673, 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. In *European Conference on Computer Vision*, pp. 405–421, 2020.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-E: A system for generating 3D point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

Liang Pan, Xinyi Chen, Zhongang Cai, Junzhe Zhang, Haiyu Zhao, Shuai Yi, and Ziwei Liu. Variational relational point completion network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8520–8529, 2021.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations*, 2023.

Qwen Team. Qwen3.8-27B. Model card, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

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

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

Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. XCube: Large-scale 3D generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4209–4219, 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.

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.

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

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *Transactions on Machine Learning Research*, 2026.

Stefan Stojanov, Anh Thai, and James M. Rehg. 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*, pp. 1798–1808, 2021.

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, 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*, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22819–22829, 2023.

Lyne P. Tchapmi, Vineet Kosaraju, S. Hamid Rezatofighi, Ian Reid, and Silvio Savarese. TopNet: Structural point cloud decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 383–392, 2019.

Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, Qingxiang Lin, Zeqiang Lai, Xianghui Yang, Huiwen Shi, Zibo Zhao, Bowen Zhang, Hongyu Yan, Lifu Wang, Sicong Liu, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Dongyuan Guo, Junlin Yu, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Shida Wei, Chao Zhang, Yonghao Tan, Yifu Sun, Lin Niu, Shirui Huang, Bojian Zheng, Shu Liu, Shilin Chen, Xiang Yuan, Xiaofeng Yang, Kai Liu, Jianchen Zhu, Peng Chen, Tian Liu, Di Wang, Yuhong Liu, Linus, Jie Jiang, Jingwei Huang, and Chunchao Guo. Hunyuan3D 2.1: From images to high-fidelity 3D assets with production-ready PBR material. *arXiv preprint arXiv:2506.15442*, 2025. URL https://arxiv.org/abs/2506.15442.

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

Hanxiao Wang, Ying-Tian Liu, Yuan-Chen Guo, Qi-Yuan Feng, Zi-Xin Zou, Ding Liang, Biao Zhang, and Yan-Pei Cao. Nexus: Native mesh generation with diffusion. *ACM Transactions on Graphics*, 45(4):71:1–71:14, 2026. doi: 10.1145/3811344.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2D diffusion models for 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12619–12629, 2023a.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89, 2025.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21686–21697, 2024a.

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

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27171–27183, 2021.

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

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

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems*, volume 38, pp. 40377–40408, 2025d. URL https://papers.nips.cc/paper_files/paper/2025/hash/336572db3e99930814d6b328d4220cb6-Abstract-Conference.html.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024b.

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. RODIN: A generative model for sculpting 3D digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4563–4573, 2023b.

Xiaogang Wang, Marcelo H. Ang, Jr., and Gim Hee Lee. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 790–799, 2020.

Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. RoboGen: Towards unleashing infinite data for automated robot learning via generative simulation. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51936–51983, 2024c. URL https://proceedings.mlr.press/v235/wang24cc.html.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8406–8441, 2023c. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1a87980b9853e84dfb295855b425c262-Abstract.html.

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. CRM: Single image to 3D textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision*, pp. 57–74, 2024d.

Xin Wen, Tianyang Li, Zhizhong Han, and Yu-Shen Liu. Point cloud completion by skip-attention network with hierarchical folding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1939–1948, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wen_Point_Cloud_Completion_by_Skip-Attention_Network_With_Hierarchical_Folding_CVPR_2020_paper.html.

Guanjun Wu, Taoran Yi, Jiemin Fang, Wenyu Liu, and Xinggang Wang. Fast high dynamic range radiance fields for dynamic scenes. In *2024 International Conference on 3D Vision (3DV)*, pp. 862–872, 2024a.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20310–20320, 2024b.

Guanjun Wu, Jiemin Fang, Chen Yang, Sikuang Li, Taoran Yi, Jia Lu, Zanwei Zhou, Jiazhong Cen, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Xinggang Wang, and Qi Tian. UniLat3D: Geometry-appearance unified latents for single-stage 3D generation. In *Findings of CVPR*, 2026. URL https://github.com/UniLat3D/UniLat3D.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125116–125141, 2024c.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3D: Scalable image-to-3D generation via 3D latent diffusion transformer. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121859–121881, 2024d.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip H. S. Torr, and Yao Yao. Direct3D-S2: Gigascale 3D generation made easy with spatial sparse attention. In *Advances in Neural Information Processing Systems*, volume 38, pp. 189214–189240, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f9666a092153f281b5116cd1f64b5c91-Abstract-Conference.html.

Jiatong Xia, Zicheng Duan, Anton van den Hengel, and Lingqiao Liu. Points-to-3D: Structure-aware 3D generation with point cloud priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19928–19939, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21469–21480, 2025.

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

Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. GRNet: Gridding residual network for dense point cloud completion. In *European Conference on Computer Vision*, pp. 365–381, 2020.

Bojun Xiong, Si-Tong Wei, Xin-Yang Zheng, Yan-Pei Cao, Zhouhui Lian, and Peng-Shuai Wang. OctFusion: Octree-based diffusion models for 3D shape generation. *Computer Graphics Forum*, 44(5):e70198, 2025.

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*, 2024.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP-2: Towards scalable multimodal pre-training for 3D understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27091–27101, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xue_ULIP-2_Towards_Scalable_Multimodal_Pre-training_for_3D_Understanding_CVPR_2024_paper.html.

Weilong Yan, Haipeng Li, Hao Xu, Nianjin Ye, Yihao Ai, Shuaicheng Liu, and Jingyu Hu. LaS-Comp: Zero-shot 3D completion with latent-spatial consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7588–7599, 2026.

Xingguang Yan, Liqiang Lin, Niloy J. Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. ShapeFormer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21924–21935, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024b.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language guided generation of 3D embodied AI environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16277–16287, 2024c.

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, 2018.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3D prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9009–9019, 2023.

Chaohui Yu, Qiang Zhou, Jingliang Li, Zhe Zhang, Zhibin Wang, and Fan Wang. Points-to-3D: Bridging the gap between sparse points and shape-controllable text-to-3D generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6841–6850, 2023. doi: 10.1145/3581783.3612232.

Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. PCN: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pp. 728–737, 2018.

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*, 42(4):1–16, 2023. doi: 10.1145/3592442.

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*, 43(4):1–20, 2024. doi: 10.1145/3658146.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Shuai Zhang, Guanjun Wu, Zhoufeng Xie, Xinggang Wang, Bin Feng, and Wenyu Liu. Dynamic 2D gaussians: Geometrically accurate radiance fields for dynamic objects. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8144–8153, 2025a. doi: 10.1145/3746027.3755328.

Yibo Zhang, Li Zhang, Rui Ma, and Nan Cao. TexVerse: A universe of 3D objects with high-resolution textures. *arXiv preprint arXiv:2508.10868*, 2025b.

Tianhao Zhao, Youjia Zhang, Hang Long, Jinshen Zhang, Wenbing Li, Yang Yang, Gongbo Zhang, Jozef Hladký, Matthias Nießner, and Wei Yang. LATO: 3D mesh flow matching with structured TOpology preserving LAtents. In *International Conference on Machine Learning*, 2026.

Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3D shape generation based on shape-image-text aligned latent representation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 73969–73982, 2023.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring unified 3D representation at scale. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/cc2c03ce028a1d6cbed6f98cede7f5a9-Abstract-Conference.html.

Linqi Zhou, Yilun Du, and Jiajun Wu. 3D shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5806–5815, 2021.

SUPRA-Vox: Native Voxel Prompts for Super-Aligned 3D Asset Generation

Supplementary Materials

A Additional Implementation Details

Training configuration. We fine-tune from the released Pixal3D checkpoint with AdamW ($\beta = (0.9, 0.95)$, weight decay 0.01), a learning rate of 10^{-4}, $1{,}000$ linear warmup steps, gradient clipping at 1.0, and bfloat16 mixed precision. The global batch size is 256 across 64 GPUs with 80 GB of memory each. Training runs for approximately 30k steps and uses nearly 700 GPU-hours. Reported weights are not averaged with an exponential moving average (EMA).

Conditioning. The image conditioner is a DINOv3 (Siméoni et al., 2026) ViT-L/16 projection feature extractor with input resolution 512 and grid resolution 16. The image condition is dropped with probability 0.1. Training prompts use depth rendered from each object's mesh at the input view, whereas inference prompts use MoGe-3 (ViT-L). Matched GT-depth evaluations diagnose sensitivity to this substitution and are reported as oracle results.

Inference. Sparse voxel generation uses a 12-step Euler solver, random seed 42, and no classifier-free guidance. At each sampling step, the surface prompt and mask are provided as conditioning inputs, and the noisy state is updated using the predicted velocity. The frozen Pixal3D Shape1024 generator and decoder produce the final mesh.

B Data and Evaluation Details

B.1 Data Filtering

Annotation and selection pipeline. Figure 6 summarizes asset annotation and training-data selection. We use a 6-view Qwen3.8-27B (Qwen Team, 2026) VLM pass for structural categorization, including isolated objects, indoor fragments, extraneous ground, and outdoor terrain. The selector consumes these structural categories together with recorded quality, complexity, and score fields. The composite score combines the raw VLM score with mesh- and texture-complexity grades, as defined below. Selection statistics are computed from the frozen manifests.

Scope and quality criteria. We report the curation of the six source collections represented in HQ100k: TexVerse (Zhang et al., 2025b), the Sketchfab and GitHub subsets of Objaverse-XL (Deitke et al., 2023), ABO (Collins et al., 2022), HSSD (Khanna et al., 2024), and Toys4K (Stojanov et al., 2021). The goal is to retain recognizable objects with useful surface structure and valid supervision for image-conditioned geometry generation.

Meaning of the annotation fields. The structural label describes the kind of asset: `normal` denotes an ordinary object. The separate `quality_level` field is an ordinal quality label (low, medium, or high), whereas `poly_level` records an ordinal geometry-complexity label (low-, mid-, or high-poly). We write $s(a)$ for the precomputed `total_score` of asset a. This composite score combines the raw VLM rating with mesh-complexity grades (low- to high-poly) and texture-complexity grades (simple to complex textures). These inputs describe complementary aspects of an asset: the VLM rating provides an overall assessment, the mesh grade describes geometric complexity, and the texture grade describes the complexity of surface appearance. The resulting score is stored in the offline annotation metadata and used for eligibility and ranking, with higher values preferred. The inherited `aesthetic_score` is a separate metadata field.

Eligibility rule. The selector retains assets labeled `normal`, with medium or high quality, mid- or high-polygon complexity, and a recorded total quality score of at least 7. Among the retained objects, $89{,}999$ are labeled `medium` and $13{,}550$ are labeled `high`. Likewise, $74{,}519$ are labeled `mid_poly` and $29{,}030$ are labeled `high_poly`.

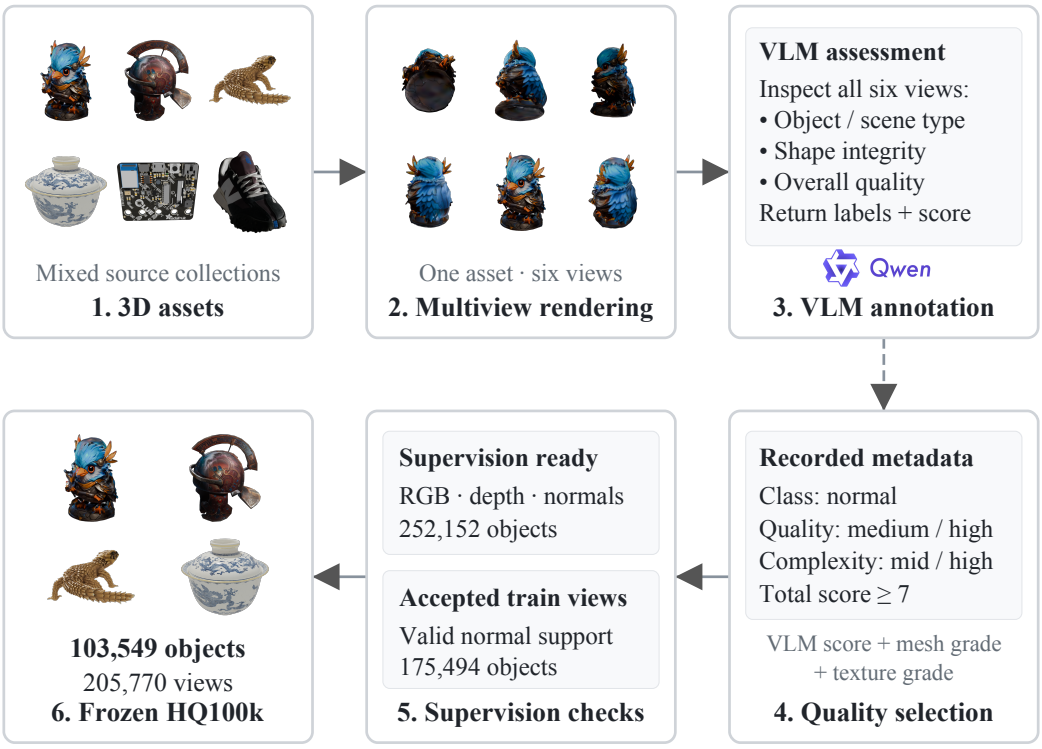


Figure 6: **Asset annotation and HQ100k selection.** Asset renders illustrate the input collection and multi-view observations. The schematic VLM assessment supplies labels and scores for offline selection, which combines the VLM score with mesh- and texture-complexity grades. Supervision checks produce the final score-ranked pool from six sources.

Preprocessing readiness. Candidates must also have completed conditional rendering, normal and geometry preprocessing, ground-truth depth, and per-view shape-latent and dual-grid preprocessing for all 16 views. Within the six reported sources, these combined quality and readiness gates yield 252,152 candidates.

Accepted views and global ranking. We next join candidates with the accepted training-view normal manifest, whose validity criterion requires at least 2,048 valid normal pixels. Excluding 3,655 objects in the normal validation split and 73,003 without accepted training or validation entries leaves 175,494 objects. The final set uses the frozen score-ranked membership, without per-source quotas. Ranking ties are resolved by a fixed dataset order and then asset identifier. The retained six-source pool has a minimum score of 7.33, distinct from the eligibility threshold of 7; the median is 7.57.

HQ100k and source composition. The resulting frozen pool contains 103,549 objects and 205,770 views: TexVerse (70,032), Sketchfab (17,175), GitHub (13,978), ABO (1,199), HSSD (922), and Toys4K (243). Source contributions follow the global quality-score ranking. Figure 7 reports the selection counts and score distributions from the frozen manifests. Of the retained objects, 1,328 have one accepted view and 102,221 have two; these accepted views are distinct from the 16 preprocessed views required by the readiness check.

Visual examples and boundary cases. Figure 8 shows four retained assets with high quality labels and four nonmember examples. Retained examples are chosen for clear shape and diverse appearance. The nonmember examples include terrain, partial room assemblies, and a low-poly asset outside the selected complexity range. Low-poly and untextured assets can still have valid geometry, reflecting the task-specific nature of selection. For display, we trim transparent borders and use a common scale within each asset's three original rendered views.

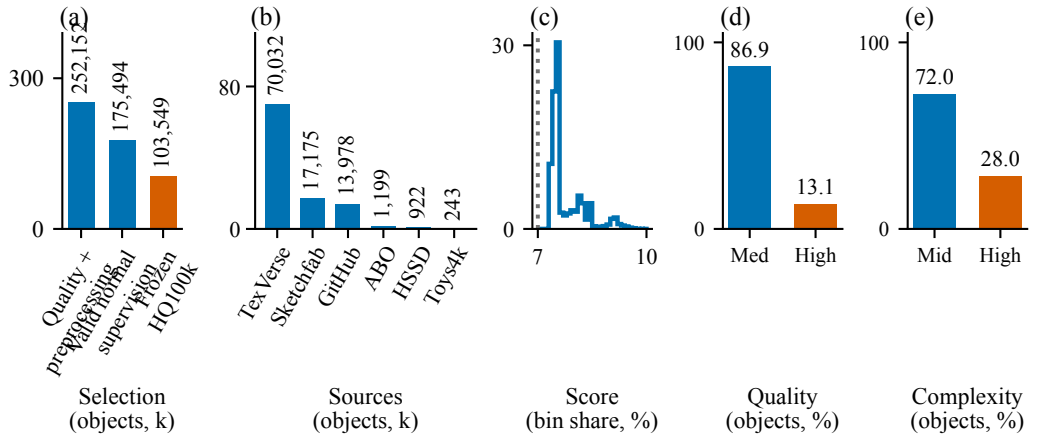


Figure 7: **HQ100k statistics for the six reported sources.** (a) Objects after quality and preprocessing checks, after requiring valid normal supervision, and in the frozen HQ100k selection. A valid normal supervision view belongs to the training split and contains at least 2,048 pixels in its explicit normal-validity mask. (b) Final source composition. In (a–b), vertical axes are in thousands of objects and the labels above bars give exact object counts. (c) Distribution of the recorded `total_score` over 103,549 selected objects: each height is the percentage in a bin of width 0.1, and the dotted line marks $s = 7$. (d) Quality-label proportions (Med: medium; High: high). (e) Complexity-label proportions (Mid: mid-poly; High: high-poly). Panels (d–e) show the proportions of each label in the selected pool.

B.2 Geometry Estimator and Depth-Condition Ablations

Settings. We evaluate two inference-time substitutions on Complex-Geo Subset, selected from our evaluation set, keeping the generator fixed. The reference uses MoGe-3, Native Voxel Prompts, and the visual hull mask. The *MoGe-2* variant uses MoGe-2 ViT-L in full precision to predict depth and camera intrinsics, then rebuilds the prompts and masks with the reference filtering, fixed calibration anchors, and canonicalization rules. The *depth-zero* variant sets both the encoded prompt and mask to zero, $(\mathbf{z}^{\mathrm{str}}_{\mathrm{obs}}, \mathbf{m}_{\mathrm{obs}}) \leftarrow (\mathbf{0}, \mathbf{0})$, at every sampling step. It retains the reference RGB input and MoGe-3-derived camera, isolating the effect of explicit depth conditioning without retraining. It therefore does not represent inference without a geometry estimator.

Evaluation. All settings share the input images and supplied foreground masks, the frozen Shape1024 generator and decoder, 12 sampling steps, and seed 42. The supplied foreground masks make this a semi-information evaluation. We report means over the selected subset using the same normal-render PSNR/SSIM and whole-mesh metrics as the main ablations. F-score and normalized Chamfer-L_1 use 10,000 surface samples and Sim(3) scoring alignment. No additional mesh postprocessing is used.

Results. MoGe-2 completes all cases without adapting the generator, but reduces normal PSNR by 0.293 dB and mean F-score by 0.0183 (Table 4). Removing the depth condition reduces normal PSNR by 1.575 dB and mean F-score by 0.0292. Paired bootstrap intervals over objects (10,000 resamples) exclude zero at the 95% level for all four depth-zero differences. For MoGe-2, the intervals exclude zero for PSNR, SSIM, and Chamfer, but include zero for F-score. These results support the contribution of explicit depth conditioning and compatibility with a second estimator, while showing that accuracy remains sensitive to the observation source.

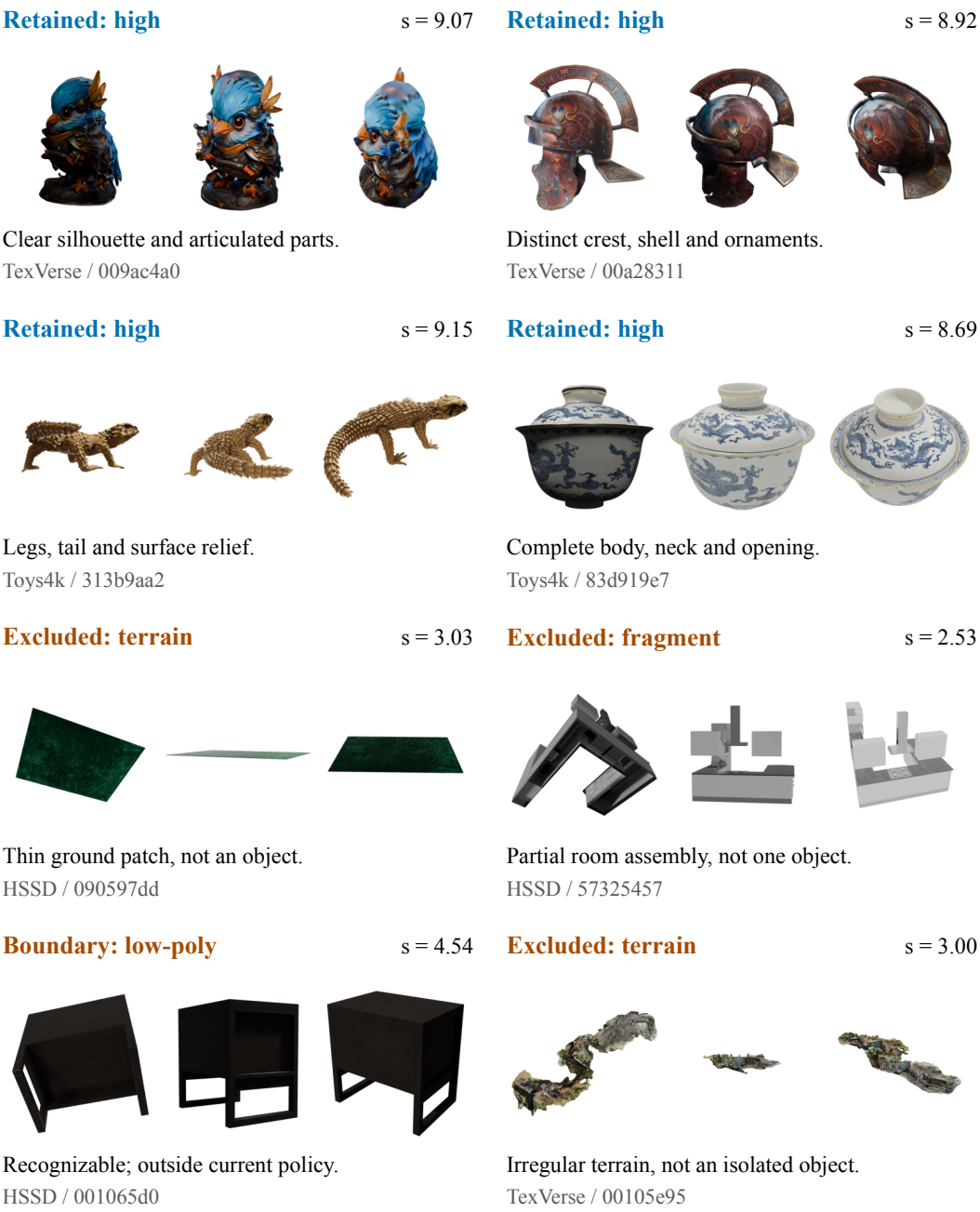


Figure 8: **Real multiview examples of selection criteria.** Each asset shows three rendered views, its composite score s, source, and identifier prefix. The first two rows show retained, high-labeled assets; the last two show terrain patches, a partial indoor assembly, and a low-poly boundary case. Scores and membership come from metadata, while the short descriptions explain the selection criteria.

B.3 Bounding-Box Calibration

We align the predicted depth surface to the visible first-hit surface of native RGB-conditioned Pixal3D sparse voxels. Ground-truth geometry is not used in this calibration. Let $\mathbf{p}$ denote the lifted points, and $\mathbf{q}$ the reference points. The fixed axis conversion is $\widetilde{\mathbf{p}} = \mathrm{diag}(1, -1, -1)\mathbf{p}$. For each point set, compute coordinate-wise 0.001 and 0.999 quantiles, giving bounds $\mathbf{a}_p, \mathbf{b}_p$ and $\mathbf{a}_q, \mathbf{b}_q$.

Write $\mathbf{e} = \mathbf{b} - \mathbf{a}$. The uniform scale and lateral translation are

$$s = \frac{e_{p,x}e_{q,x} + e_{p,y}e_{q,y}}{e_{p,x}^2 + e_{p,y}^2}, \qquad \mathbf{t}_{xy} = \frac{(\mathbf{a}_q + \mathbf{b}_q)_{xy}}{2} - s\frac{(\mathbf{a}_p + \mathbf{b}_p)_{xy}}{2}. \tag{9}$$

For points paired by input pixel, form $r_i = q_{i,z} - s\widetilde{p}_{i,z}$ and set

$$t_z = \frac{Q_{0.25}(r) + Q_{0.50}(r) + Q_{0.75}(r)}{3}, \qquad \mathbf{p}_i^{\text{can}} = s\widetilde{\mathbf{p}}_i + (\mathbf{t}_{xy}, t_z). \tag{10}$$

Here Q_α is the empirical quantile. The camera is transformed consistently with the points. Lifting and image conditioning share the selected intrinsics. The front-surface bounding-box ablation maps the predicted surface's $0.001/0.999$ quantile axis-aligned bounding box (AABB) directly to the canonical cube.

B.4 Evaluation Dataset Construction

We construct two complementary object collections for evaluating image-conditioned geometry generation. Complex-Geo contains 115 assets curated from two collections of geometrically detailed models, after duplicate removal and manual exclusions. Its examples include architecture, mechanical assemblies, branching structures, and objects with open lattice surfaces. DTC-Geo contains 100 assets selected from DTC using a geometry-complexity ranking, after removing overlaps with the Complex-Geo source collections and manually excluded candidates. The highest-ranked remaining 100 assets form the evaluation set. Examples include household objects and ornamental models with intricate contours and surface relief.

Each object is associated with a ground-truth mesh normalized to the canonical box $[-0.5, 0.5]^3$, 16 rendered views and their camera parameters. Evaluation uses one designated reference view recorded in the dataset manifest; DTC-Geo reference views are selected manually. The reference image conditions generation, while ground-truth geometry and rendered normals provide evaluation targets. Figure 9 shows eight selected examples from each collection to highlight shape diversity.

B.5 Evaluation Protocol

Visible-surface evaluation. Complex-Geo and DTC-Geo use their designated reference views. Normals are rendered at 1024^2 in the camera frame, encoded as $(\mathbf{n} + 1)/2$, and evaluated at 512^2. For full-image PSNR, SSIM, and VGG-LPIPS, invalid pixels in each image are independently filled with 0.5. PSNR uses peak value 1, SSIM uses a 7×7 window, and LPIPS receives RGB mapped to $[-1, 1]$. Silhouette IoU measures mask intersection over union. Normal angular errors are measured only on the mask intersection after unit normalization. We report mean and median errors, strict angular accuracies below $11.25°$, $22.5°$, and $30°$, and mean boundary error. The boundary is a radius-8 disk dilation-minus-erosion band of the overlap mask, restricted to the overlap at evaluation resolution. Pixal3D and SUPRA-Vox renderings are placed back on the reference canvas using the same $1.1\times$ object-box uncrop convention. Native Pixal3D uses the designated field of view with camera distance $d = 1/(2\tan(\text{fov}/2))$ and identity object rotation. SUPRA-Vox and its GT-depth control use their inference cameras with the same reference-canvas placement. Evaluation uses the known reference cameras.

Whole-mesh alignment and geometric metrics. The geometric columns in Table 2 evaluate the 115 Complex-Geo objects. We sample 10,000 surface points per mesh with seed 42. View-based camera alignment is followed by centroid and RMS-scale alignment and 30 untrimmed Sim(3) iterative closest point (ICP) iterations, allowing rotation, translation, and unconstrained uniform scale to change. These metrics assess shape after ground-truth-assisted alignment. SAM 3D uses its official postprocessed layout when accepted ($47/115$ objects), and its metadata layout otherwise. Chamfer L_1 averages the bidirectional coordinate-wise L_1 distances using Euclidean nearest-neighbor correspondences. F-score uses Euclidean distance threshold 0.01 and is reported as a percentage. EMD uses 1,024 points, uniform masses, and Euclidean transport cost; the displayed value is Wasserstein-1 distance multiplied by 100. The vIoU column measures occupancy IoU of the sampled surface points on a 64^3 grid in $[-0.5, 0.5]^3$, clipping out-of-range indices to the grid.



Figure 9: **Snapshot of the evaluation datasets.** Eight examples from Complex-Geo (115 objects) and eight from DTC-Geo (100 objects). Images are ground-truth asset renders from the designated conditioning views, with identifier prefixes shown below. The gallery illustrates multi-part shapes, thin structures, openings, and surface detail. Objects are rescaled individually for display.

Image–3D agreement. For the Uni3D and ULIP-2 columns, we sample 10,000 mesh vertices, padding by resampling when needed, with seed 42. Points have constant RGB 0.4 and are centered and normalized to a unit sphere by maximum radius. Uni3D-G and ULIP-2 Colored PointBERT embeddings are compared with four ground-truth rendered views at yaw $\{-16, 74, 164, 254\}^\circ$ and pitch 20°. We average normalized embedding cosine similarities over these views and multiply by 100. Missing meshes and scoring failures receive zero in this protocol.

Observation controls. Matched MoGe and GT-depth observations are used for the main comparisons. The GT-depth variant is an oracle observation control.

B.6 In-the-Wild Evaluation and User Study

We evaluate image–3D agreement on a 50-case in-the-wild image set, without ground-truth cameras or normals. Following the protocol in Table 2 of Pixal3D (Li et al., 2026a), Uni3D and ULIP-2

compare each generated shape with its single conditioning image using 8,192 area-weighted surface samples. Table 3 reports the available results.

Video-based user study. Ten experienced evaluators each assess 15 cases, giving 150 pairwise comparisons. This mixed-source case set is separate from the 50-case automatic evaluation. Each pair contains our result and one baseline, shown as synchronized local-orbit videos beside the input image. The videos span yaw $\pm 20^\circ$ and pitch $\pm 8^\circ$. Method names are hidden, and evaluators choose A, B, or a tie for visible-surface alignment, structural plausibility, and overall quality. The main table reports the first two criteria as $100W/(W + T + L)$, where W, T, and L count our wins, ties, and losses. Hunyuan3D 2.1 and TRELLIS.2 each receive 38 comparisons; Pixal3D and UniLat3D each receive 37. The left–right presentation order of our result and the baseline is randomized.

C Additional Qualitative Comparisons

Figure 10 provides eight additional examples comparing SUPRA-Vox with UniLat3D, Hunyuan3D 2.1, TRELLIS.2, and Pixal3D.

D Failure Case: Incorrect Depth-Derived Surface Structure

Figure 11 illustrates a limitation of surface prompting. The input depicts a large front-facing wheel with pronounced radial blades. The MoGe-3 reconstruction instead represents much of this region as a smooth disk, losing the surface relief suggested by the image. The SUPRA-Vox result retains this disk-like structure, whereas native Pixal3D produces more distinct radial blades in the same region. This qualitative correspondence is consistent with an erroneous surface prior propagating into the generated geometry. Since Native Voxel Prompts guide early spatial support, an incorrect observed surface can bias the structure on which subsequent detail is synthesized; the downstream generator need not correct that error. Observation reliability therefore matters for major visible structures as well as fine detail, motivating confidence-aware prompting and improved geometry estimation.

E Field of View in Monocular Lifting

Changing the assumed field of view changes the back-projected surface even when the depth map is held fixed. Figure 5 uses one RGB image and one estimated depth map. Only the intrinsics used to lift that map change.

In this example, the 30° surface has a larger relative front-to-back extent than the 60° surface, as shown by the matched side views in Fig. 5a. This comparison illustrates sensitivity to the lifting intrinsics, rather than an accuracy ranking of the two FOV choices. The pipeline shares the selected intrinsics between lifting and image conditioning (Appendix B.3).



Figure 10: **Additional qualitative comparisons** on evaluation-set and in-the-wild images. Ours denotes SUPRA-Vox; red boxes highlight local geometry.

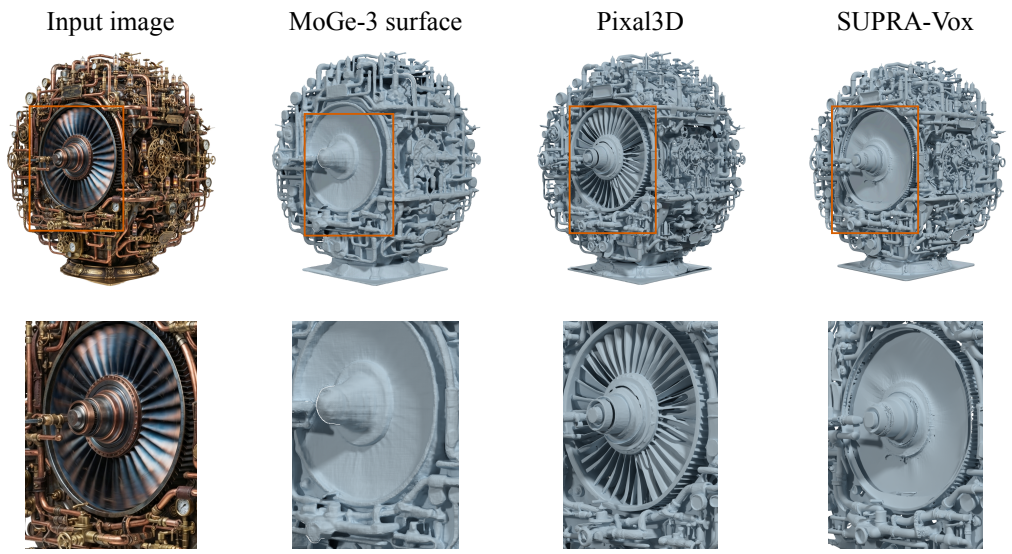


Figure 11: **Incorrect surface evidence can persist in generation.** The wheel-region enlargements show that the smooth disk-like surface in the depth reconstruction is also present in the prompted result.