COARSE3D: CONTROLLABLE AND FAITHFUL SPARSE STRUCTURE GENERATION FOR IMAGE-TO-3D

Anonymous authors
Paper under double-blind review

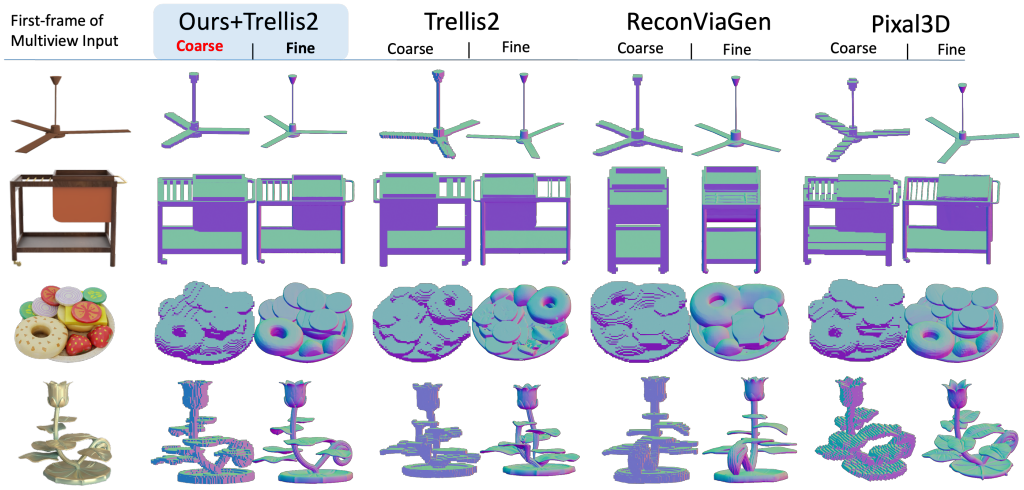


Figure 1: **Coarse3D** generates orientation-controllable and structurally faithful sparse structures. Trellis2, ReconViaGen and Pixal3D use their own coarse and refinement stages while **Ours+Trellis2** replaces *only the coarse stage* with Coarse3D and reuses the Trellis2 refinement stage. Each cell shows the coarse voxel structure (*left*) and the final shape it yields (*right*). Coarse3D consistently follows the first-view orientation and preserves structural fidelity across diverse objects, directly translating into higher-quality final shapes through the coarse-to-fine pipeline.

ABSTRACT

Coarse-to-fine 3D-native generation relies heavily on the first-stage sparse structure, which determines the global orientation and overall geometry for subsequent refinement. However, existing coarse generators suffer from two key limitations: their orientation is difficult to control due to ambiguous semantic canonicalization, and their structures are often inconsistent with input images due to weak 2D-3D grounding. We propose **Coarse3D**, a sparse-structure generation model that produces orientation-controllable and structurally faithful voxel geometry from arbitrary unposed images. Coarse3D introduces three designs: a first-frame anchored canonical space that aligns generation with the first input view while preserving object-centric normalization; a dual-tower DiT that couples generation with auxiliary feed-forward reconstruction to learn spatially grounded features; and spatial-prior guided attention that regularizes cross-attention with pixel-to-voxel correspondence. Coarse3D can directly replace the coarse stage of existing pipelines. Experiments show that it substantially improves sparse structure quality and consistently enhances final 3D generation, achieving state-of-the-art performance.

1 INTRODUCTION

The automatic creation of high-quality 3D assets from images is a long-standing goal in computer graphics and vision, with applications in gaming, AR/VR, robotics, and digital manufacturing. Recent 3D-native generative models, such as CraftsMan3D Li et al. (2025a), TRELLIS Xiang et al.

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.

(2025; 2026), Sam3D Chen et al. (2026), have made substantial progress by adopting a coarse-to-fine paradigm: a sparse geometric structure is first generated as voxels, and then refined into detailed geometry and texture. In this pipeline, the first-stage voxel structure is particularly important: it determines the global orientation and coarse geometry on which all later refinement is built, and any error introduced here is difficult to correct in subsequent stages. Yet current coarse-stage generation still suffers from two key limitations: uncontrollable orientation and unfaithful structure. In this work, we therefore focus on sparse-structure generation and propose Coarse3D, a model that produces orientation-controllable and structurally faithful voxel geometry from unposed images.

Uncontrollable voxel orientation. Much of this uncertainty stems from the object-centric canonical space used by native 3D generators (Xiang et al., 2025; Ye et al., 2025; Li et al., 2025c), where each object must be normalized to a shared frame with its semantic parts aligned to predefined axes. Recovering this frame from images alone is ill-posed: for many objects a "front" is simply undefined (e.g., the ceiling fan in Fig. 1). Even where a front is well defined, the supervision itself is inconsistent: large-scale datasets such as Objaverse-XL (Deitke et al., 2024) often store same-category assets in different orientations. The model is thus asked to fit a mapping that is neither well defined nor consistently supervised, leaving the orientation of generated assets uncontrollable.

Coarse3D instead adopts a first-frame anchored canonical space: the object is still centered at the origin and normalized to a unit cube, but its orientation is defined by the first input view rather than by a semantic front. This keeps generation in a compact, scale-normalized, camera-independent domain while eliminating the ill-posed rotation inference altogether. Note that this differs from camera-centric (Wang et al., 2025a) or pixel-aligned prediction (Li et al., 2026), where geometry is tied to the imaging frame and its position and scale depend on focal length and camera distance. Here only the orientation is borrowed from the view, and the representation itself remains object-centric.

Unfaithful voxel structure. The second limitation concerns faithfulness: as shown in Fig. 1, existing coarse generators may omit parts that are clearly visible in the inputs (the missing leaves and branches of the flower), hallucinate geometry that appears in none of the views (a spurious fifth crossbar on the cart), or misplace structures relative to the observations (the items on the plate are rearranged). We attribute this to weak 2D-3D grounding during image conditioning. At the feature level, the image features used for cross-attention encode appearance and semantics but few explicit geometric cues, leaving voxel tokens with little spatial evidence to align with and no cross-view geometry to reconcile them. At the attention level, the 2D-3D mapping is left entirely implicit: each voxel token attends to all image tokens with no geometric constraint, so pixel-to-voxel correspondence must be learned from scratch with limited 3D supervision.

To improve faithfulness, Coarse3D strengthens grounding at both levels. At the feature level, we introduce a dual-tower DiT that couples sparse-structure generation with an auxiliary reconstruction tower, injecting features layer by layer and trained end-to-end. The reconstruction objective forces the conditioning features to encode spatial and cross-view geometry and injects them into generation at every layer, avoiding the feature mismatch and information bottleneck of staged reconstruction-then-generation pipelines such as ReconViaGen (Chang et al., 2026) and RecGen3D (Huang et al., 2026). At the attention level, we propose spatial-prior-guided attention, which rasterizes the ground-truth structure into visible pixel-to-voxel associations and uses them to regularize cross-attention, encouraging each voxel token to attend to the image regions it actually projects to. As this prior is applied only during training, Coarse3D needs no camera poses at inference.

In summary, our contributions are:

- We present **Coarse3D**, a sparse-structure generator that addresses two overlooked failure modes of the coarse stage—uncontrollable orientation and unfaithful structures. As experiments demonstrate, it improves coarse-to-fine baselines on benchmarks by a large margin.
- We propose a **first-frame anchored canonical space** that anchors the generated orientation to the first input view instead of an ill-defined semantic front, preserving an object-centric unit cube representation while also accommodating generation and reconstruction in a single space.
- We strengthen 2D-3D grounding with a **dual-tower DiT** that makes conditioning features geometry-aware via auxiliary reconstruction, and a **spatial-prior guided attention** that regularizes cross-attention with rasterized pixel-to-voxel associations.

2 Related Work

2.1 3D-Native Generation

Early image-to-3D methods rely on slow per-instance optimization (Mildenhall et al., 2021; Kerbl et al., 2023; Poole et al., 2023; Tang et al., 2024). Recent 3D-native models instead generate directly in 3D latent spaces. One family encodes shapes as unstructured vector sets (Zhang et al., 2023; 2024; Team, 2025; Ye et al., 2025) and decodes geometry in a single step, leaving no intermediate structure to control. The other, introduced by TRELLIS (Xiang et al., 2025), attaches structured latents to active voxels of a sparse grid, splitting generation into a coarse-to-fine cascade: a first stage produces the sparse structure, a second predicts per-voxel latents, and a decoder yields the shape. TRELLIS2 (Xiang et al., 2026) extends it with field-free O-voxels and separate shape and material stages and several other TRELLIS-based variants have also emerged (Wu et al., 2025; Li et al., 2025d; He et al., 2025; Zhao et al., 2026). Because the cascade is sequential, the first stage fixes global orientation and layout and upper-bounds later stages, yet prior work mainly targets representation design and shape refinement. We instead focus on this coarse stage, identifying orientation controllability and structural faithfulness as its two bottlenecks.

2.2 Orientation in 3D Generation

Most 3D-native generators place objects in an object-centric canonical space (Zhang et al., 2024; Xiang et al., 2025; Li et al., 2025b): shapes are normalized to a centered, unit-scale domain, and their semantic parts are expected to align with predefined axes. The normalization itself is convenient, but the axis-alignment requirement forces the model to infer a semantic orientation from the input images, e.g., which side is the front. This inference is ill-posed for objects without a clear front and is further confounded by the inconsistent orientations of assets in large-scale datasets (Deitke et al., 2024; Zhang et al., 2025). Prior work attacks this from two directions. OrientationMatters (Lu et al., 2025) reduces dataset-level inconsistency by curating an orientation-aligned dataset with VLM-based front-view prediction and manual calibration, which is costly and still presumes a well-defined semantic front. Pixal3D (Li et al., 2026) instead sidesteps the inference by defining geometry in the input camera frame, so the output aligns with the input view, but this leaves the shared canonical domain and makes the output dependent on camera parameters and depth scale. Coarse3D keeps the centered, unit-scale canonical domain and anchors only the orientation to the first input view, thus removing semantic-front inference without giving up canonical normalization.

2.3 Reconstruction Grounding and 2D-3D Correspondence

Structural faithfulness hinges on how well image evidence is grounded in 3D. Most 3D-native generators condition voxel tokens on image features via cross-attention (Xiang et al., 2025; Li et al., 2025b; Fang et al., 2026), leaving the 2D-3D relation entirely implicit. Reconstruction priors provide stronger geometric cues: ReconViaGen (Chang et al., 2026) injects multi-view stereo features from VGGT (Wang et al., 2025a) into a TRELLIS-style generator, and RecGen3D (Huang et al., 2026) uses canonicalized reconstructed point clouds as geometric anchors for diffusion. Both, however, keep reconstruction and generation in separate stages and never in a common frame: ReconViaGen leaves reconstruction in a camera-centric frame while generating in an object-centric canonical frame, and RecGen3D bridges the two only by explicitly canonicalizing the reconstructed point cloud. Coarse3D instead defines a single first-frame anchored space for both, and places feed-forward reconstruction in an auxiliary tower of the same DiT, sharing features layer by layer so that the generation tower consumes reconstruction-grounded features directly and end-to-end.

Even with geometry-aware features, cross-attention must still discover pixel-to-voxel correspondence on its own, which is difficult at the scale of available 3D data. ROAR-3D (Sun et al., 2026) routes each voxel token to the most relevant views, but the resulting alignment is view-level and does not indicate where inside a view a token should attend. Pixal3D (Li et al., 2026) instead builds pixel-aligned features by back-projection in camera space, which requires camera parameters at inference. Coarse3D derives pixel-to-voxel associations by rasterization and uses them purely as a training-time regularizer in spatial-prior guided attention, encouraging geometrically valid alignment without modifying the attention formulation or requiring camera parameters at test time.

3 METHODOLOGY

3.1 FIRST-FRAME ANCHORED CANONICAL SPACE

Existing coordinate systems face a dilemma (Fig. 2): a standard canonical space fixes the axes but requires inferring an ambiguous semantic front (a teapot's spout or its handle?), whereas a pixel-aligned camera space removes this ambiguity by tying geometry to the imaging frame, making position and scale camera-dependent. Our first-frame anchored canonical space bridges the two: the object stays centered in a unit cube and we anchor only the azimuth to the first view, leaving the canonical up-axis intact, which turns the ill-posed front inference into a well-posed alignment task.

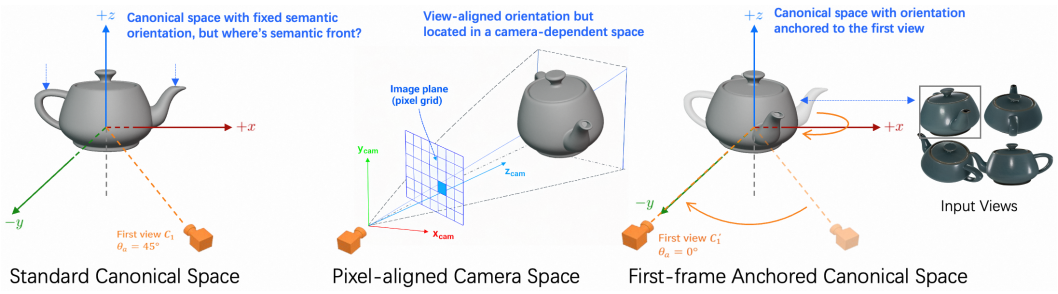


Figure 2: Comparison of three coordinate definitions: canonical space uses a fixed semantic orientation that is ambiguous; pixel-aligned camera space is aligned with the input view but not object-centric; ours is aligned with the first input view and object-centric.

Anchoring the orientation of generation. Given N input views, we take the first view as the orientation reference. Let θ_a be its azimuth in the standard canonical space, measured with respect to the $-y$ direction. To align the object orientation with the first input view, we rotate it around the canonical z-axis by an anchoring rotation $\mathbf{A} = \mathbf{R}_z(-\theta_a)$. This exactly cancels the reference view's azimuth. To keep the rendered images valid without re-rendering, we apply the inverse rotation to the cameras: with world-to-camera extrinsics $\mathbf{M}_i = [\mathbf{R}_i \mid \mathbf{t}_i]$, the updated extrinsics become

$$\mathbf{M}'_i = [\,\mathbf{R}_i\mathbf{A}^\top \mid \mathbf{t}_i\,], \qquad i = 1, \ldots, N. \tag{1}$$

Since $\mathbf{A}$ rotates about the origin and the object is centered there, the translations remain unchanged. Note that θ_a is only used for training-time preprocessing. At inference the anchoring is implicit in the learned mapping, so no camera parameters are needed.

Redefining reconstruction targets. VGGT (Wang et al., 2025a) predicts geometry in the coordinate frame of the first camera; our space keeps the same reference view but stays object-centric, with a fixed origin, unit scale, and unambiguous orientation. We can therefore recast reconstruction in exactly the frame used for generation: we supervise the camera head with the updated extrinsics $\mathbf{M}'_i$ and the point head with the anchored point maps $\mathbf{P}_i = \mathbf{A}\mathbf{P}_i^{\text{orig}}$, where $\mathbf{P}_i^{\text{orig}}$ is expressed in the standard canonical space. This unifies generation and reconstruction in a single coordinate frame and provides the geometric foundation for the dual-tower architecture introduced next.

3.2 RECONSTRUCTION-GROUNDED DUAL-DIT

To ground generation in visual evidence, our Dual-DiT couples the generative backbone with an auxiliary feed-forward reconstruction branch (Fig. 3 (B)): the *image tower* performs reconstruction, while the *geometry tower* performs iterative denoising. The two towers have matched depth, and the geometry tower cross-attends to the image tower at every layer in the shared anchored frame, so features shaped by the reconstruction objective ground generation hierarchically.

View-aware image conditioning. Given N unposed input views, we extract per-view patch tokens using a frozen DINO encoder (Siméoni et al., 2025) and concatenate them to form a single conditioning sequence $\mathbf{C}$. Since plain concatenation hides which view a token comes from, we add a learnable view embedding $\mathbf{e}_i$ to the token set $\mathbf{C}_i$ of the i-th view, $\mathbf{C}_i \leftarrow \mathbf{C}_i + \mathbf{e}_i$. The $\mathbf{e}_i$ are initialized near zero to avoid perturbing the pretrained features early in training.

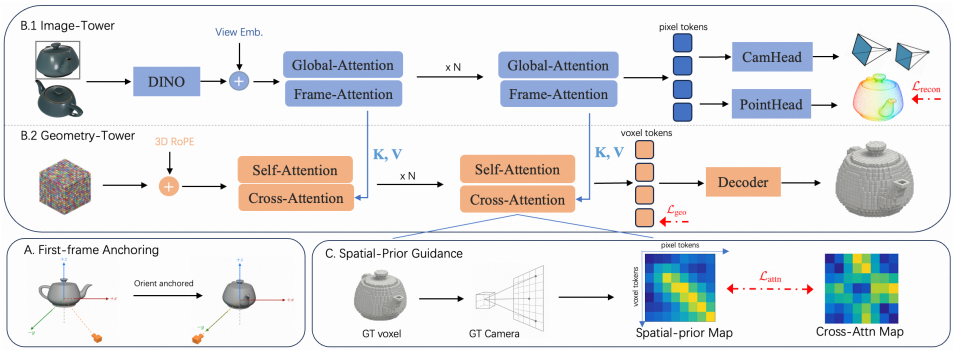


Figure 3: **Overview of Coarse3D. (A) First-frame Anchoring** rotates the canonical geometry to zero the first view's azimuth, turning orientation prediction into a well-defined, first-frame-conditioned task. **(B) Dual-tower DiT** couples an Image-Tower, supervised by auxiliary tasks ($\mathcal{L}_{\text{recon}}$) to encode cross-view geometry, with a Geometry-Tower predicting the velocity field ($\mathcal{L}_{\text{geo}}$) via layer-wise cross-attention, trained end-to-end. **(C) Spatial-prior-guided Attention** regularizes cross-attention with a rasterized voxel-to-pixel prior ($\mathcal{L}_{\text{attn}}$), strengthening 2D-3D correspondence.

Image tower. Following the VGGT architecture (Wang et al., 2025a), the image tower builds multi-view stereo associations via interleaved frame-wise and global self-attention, producing features $\mathbf{Z}^{(\ell)}$ at every layer ℓ. For efficiency we retain only the camera and point heads, attached to the top-layer features $\mathbf{Z}^{(L)}$ and supervised by the ground-truth anchored extrinsics $\mathbf{M}'_i$ and point maps $\mathbf{P}_i$. Beyond these predictions, all intermediate $\mathbf{Z}^{(\ell)}$ serve as hierarchical conditioning for generation.

Geometry tower. The geometry tower denoises the structure latent tokens $\mathbf{H}^{(\ell)}$: each layer applies self-attention over the latents, cross-attention to $\mathbf{Z}^{(\ell)}$ from the corresponding image-tower layer, and an FFN. Consuming layer-matched features avoids both the information bottleneck of compressing conditioning into a few tokens and the domain gap of staged feature injection. The final latent $\mathbf{H}^{(L)}$ is projected to the flow-matching velocity $\hat{\mathbf{v}}$.

Training objective. Both towers are trained jointly with

$$\mathcal{L} = \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{recon}}, \quad \mathcal{L}_{\text{recon}} = \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_{\text{pmap}}\mathcal{L}_{\text{pmap}}, \quad \mathcal{L}_{\text{geo}} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left\|\hat{\mathbf{v}} - (\epsilon - \mathbf{x}_0)\right\|_2^2 \tag{2}$$

where $\mathcal{L}_{\text{geo}}$ is the conditional flow-matching loss (Lipman et al., 2023), and $\mathcal{L}_{\text{cam}}$ and $\mathcal{L}_{\text{pmap}}$ are the camera and confidence-aware point-map losses of VGGT (Wang et al., 2024; 2025a). The reconstruction terms shape the image-tower features toward spatial grounding, while $\mathcal{L}_{\text{geo}}$ back-propagates through the cross-attention into the same tower, so one set of features must serve both objectives.

3.3 Spatial-Prior Guided Attention

While the Dual-DiT provides feature-level grounding, its cross-attention still has to discover pixel-to-voxel correspondence implicitly. We therefore add a short fine-tuning stage that casts the geometric projection as a spatial prior and blends it with the model's own attention to form a pseudo ground-truth target (Fig. 3 (C)). The Dual-DiT of Sec. 3.2 serves as the base model, and a frozen copy of it supplies the reference attention. Let V denote the number of voxel tokens, U the number of image tokens, and H the number of cross-attention heads.

Spatial attention prior. Rasterizing the ground-truth structure in the first-frame anchored canonical space assigns each image token $u \in \{1, \dots, U\}$ the position $\mathbf{p}_u \in [0, 1)^3$ of the visible surface voxel it covers. Since $\mathbf{p}_u$ and the voxel centers $\mathbf{c}_i$, $i \in \{1, \dots, V\}$, live in the same frame, they can be compared directly, giving a truncated Gaussian prior $\mathbf{A}^{\text{sp}} \in \mathbb{R}^{V \times U}$:

$$A^{sp}_{i,u} = \exp\left(-\frac{\|c_i - p_u\|_2^2}{2\delta^2}\right) \cdot \mathbf{1}\{\|c_i - p_u\|_2 \leqslant \delta\} \tag{3}$$

where δ is a local distance threshold. Each row of $\mathbf{A}^{\text{sp}}$ is normalized to sum to one; rows with empty support, i.e. voxels invisible from all views, are excluded from the loss.

Pseudo ground truth. Supervising the attention with the spatial prior alone would override the semantic associations that the base model has already learned. We therefore build the target on top of the model's own attention: we read the dense attention tensor $\mathbf{A}^{\text{dense}} \in \mathbb{R}^{H\times V\times U}$ from the frozen base model on the same input, and superimpose the spatial prior on it to form a pseudo ground-truth target $\mathbf{A}^{\text{gt}} \in \mathbb{R}^{H\times V\times U}$:

$$\mathbf{A}^{\text{gt}} = (1-\beta)\mathbf{A}^{\text{dense}} + \beta\mathbf{A}^{\text{sp}}, \tag{4}$$

where $\beta \in [0,1]$ controls how much geometry is injected. Since each head's target is anchored at its own distribution, the supervision adds geometric grounding on top of the existing semantics rather than pulling all heads toward one shared pattern.

Fine-tuning objective. We steer the predicted cross-attention $\mathbf{A}^{\text{pred}} \in \mathbb{R}^{H\times V\times U}$ toward this pseudo ground truth with a Kullback–Leibler (KL) divergence loss:

$$\mathcal{L}_{\text{attn}} = \frac{1}{HV}\sum_{h=1}^{H}\sum_{i=1}^{V}\text{KL}\Big(\mathbf{A}^{\text{gt}}_{h,i} \,\|\, \mathbf{A}^{\text{pred}}_{h,i}\Big), \tag{5}$$

where $\mathbf{A}^{\text{gt}}_{h,i}$ and $\mathbf{A}^{\text{pred}}_{h,i}$ are the distributions over image tokens for voxel i at head h. We fine-tune the base model with $\mathcal{L}_{\text{total}} = \mathcal{L} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}}$, where $\mathcal{L}$ is defined in Sec. 3.2.

4 Experiments

4.1 Experimental Setup

Datasets. We train Coarse3D on about 2M 3D assets: TRELLIS-500K (Xiang et al., 2025), ∼500K filtered assets from TexVerse (Zhang et al., 2025), and ∼1M *orientation-unaligned* in-house assets covering commodities, furniture, and digital humans, all processed with the TRELLIS pipeline. We evaluate on Toys4K (Stojanov et al., 2021) and Google Scanned Objects (GSO) (Downs et al., 2022).

Implementation details. The sparse-structure (SS) stage is a single Dual-DiT of about 2.0B parameters, whose geometry and image towers are initialized from the TRELLIS2 SS-DiT and VGGT (Wang et al., 2025a). The structure resolution is set to 64. Each iteration samples $1 \sim 6$ unposed views. For spatial-prior-guided attention we use $\delta = 0.1$, $\beta = 0.2$ and $\lambda_{\text{attn}} = 0.1$. We train the base model for 200K steps at a learning rate of 1×10^{-4}, then fine-tune it with the prior for 100K steps at 1×10^{-5}, with batch size 256 on 64 H20 GPUs. See Appendix A.2 for full details.

Baselines and integration protocol. We compare against the multi-view (MV) variants of TRELLIS (Xiang et al., 2025), TRELLIS2 (Xiang et al., 2026), ReconViaGen (Chang et al., 2026), and Pixal3D (Li et al., 2026). Coarse3D shares the coarse-stage interface of the first three, so we replace *only* that stage with Coarse3D and keep their refinement models frozen, isolating the effect of the coarse stage. In contrast, we do not apply Coarse3D to Pixal3D for two reasons: (i) it operates in a pixel-aligned space rather than a canonical one, and (ii) its multi-view variant relies on accurate camera poses to align cross-view features, whereas our setting assumes unposed views (detailed in Appendix A.5). Following the evaluation protocol of ReconViaGen, all methods receive 4 arbitrary unposed views, except that Pixal3D's multi-view variant additionally requires camera poses. We therefore report its native results, obtained with ground-truth poses, as an extra reference.

Evaluation metrics. We report three groups of metrics. **(1) Coarse structure:** IoU and PCA-aligned IoU (PA-IoU), whose gap indicates how much error stems from orientation, and Chamfer Distance (CD) between voxel centers. **(2) Fine shape:** CD, F-Score, and semantic consistency metrics via ULIP-2 (Xue et al., 2024) and Uni3D (Zhou et al., 2024). **(3) Auxiliary reconstruction:** camera RRE/RTE and point-map CD in the anchored canonical space. Note that all CD metrics above are computed after rigid alignment. Detailed explanations of metrics are in Appendix A.3.

4.2 Comparison

4.2.1 Coarse Structure Generation

Each method in Table 1 is compared against the ground truth in the frame it is trained to predict: canonical for Trellis, aligned for Pixal3D, and anchored for Coarse3D. IoU therefore entangles orientation and structure, while PA-IoU and CD factor out orientation and indicate structural accuracy.

Table 1: Coarse structure generation on GSO and Toys4K. IoU is orientation-sensitive, PA-IoU factors out global orientation, and CD is computed after rigid alignment. Note that Pixal3D-MV* relies on ground-truth camera poses to align and fuse features across the multiple input views, whereas the other methods operate on fully unposed images. Best in bold.

	GSO			Toys4K		
Method	IoU ↑	PA-IoU ↑	CD ↓	IoU ↑	PA-IoU ↑	CD ↓
TRELLIS-MV	0.2243	0.2909	0.0320	0.3516	0.3785	0.0246
TRELLIS2-MV	0.3098	0.3655	0.0274	0.3840	0.4412	0.0221
ReconViaGen	0.2323	0.2984	0.0293	0.2818	0.3228	0.0271
Pixal3D-MV*	0.4419	0.4420	0.0181	0.5079	0.5080	0.0133
Coarse3D (Ours)	**0.5101**	**0.5364**	**0.0124**	**0.6315**	**0.6368**	**0.0091**

Structural fidelity. Coarse3D attains the best PA-IoU (0.5364 on GSO and 0.6368 on Toys4K) and the lowest CD (0.0124 and 0.0091), i.e. about $1.45\times$ the PA-IoU and less than half the CD of the strongest unposed baseline. As both metrics are orientation-free, these gains indicate that Coarse3D produces intrinsically more accurate sparse structures rather than merely better-posed ones.

Orientation controllability. Baselines exhibit an IoU vs PA-IoU gap of up to 30% of their IoU, showing that a substantial part of their error comes from incorrect orientation. Coarse3D reduces this gap to within 0.03 (below 6%). Its gain over the strongest unposed baseline is thus larger in IoU ($1.6\times$) than in PA-IoU ($1.45\times$). The surplus is attributable to correct orientation alone.

Qualitative results. Fig. 1 shows the same trend: the baselines often settle on an unintended global orientation, or produce structural artifacts such as missing parts, hallucinated geometry, and confused repetitive components. Coarse3D instead follows the orientation of the first input view and faithfully preserves the observed geometry.

4.2.2 Fine-Grained Shape Generation with Drop-in Replacement

Analysis. Table 2 reports final shape quality when only the coarse stage of each baseline is replaced by Coarse3D, denoted by $+$ **Ours**. The replacement improves every metric on both datasets and for all three pipelines: CD drops by $3 \sim 6\times$ (e.g. $0.0613 \rightarrow 0.0127$ for TRELLIS2-MV on GSO) and F-Score rises by $0.24 \sim 0.44$. Notably, the three variants become almost indistinguishable in geometry (GSO CD within $[0.0127, 0.0132]$) despite their different refinement stages, indicating that the coarse structure rather than the refinement stage dominates the final geometry metrics. The semantic scores improve more modestly (e.g. Uni3D $0.2707 \rightarrow 0.2836$ for ReconViaGen), consistent with their limited sensitivity at this level of shape quality. As CD and F-Score are computed after rigid alignment, these gains reflect structure rather than pose, and Fig. 4 shows where they come from: repeated parts that baselines fuse or duplicate (e.g., the rack's grid shelves), thin structures they break (e.g., the sandal's spiral straps), and salient parts they hallucinate or omit (e.g., the ship's masts and sails) are all recovered after replacing the first stage with Coarse3D.

4.2.3 Auxiliary Reconstruction Task

Analysis. Table 3 checks whether the image tower really acquires reconstruction-grade geometry. Although reconstruction is only an auxiliary objective, Coarse3D is the best on every metric: it cuts the rotation error of the reconstruction foundation VGGT by about $5\times$ ($10.54 \rightarrow 2.02$ on GSO) and is on par with the specialized RecGen3D*, showing that end-to-end training with generation objective does not sacrifice reconstruction accuracy.

4.3 Ablation Study

We ablate the three designs by adding them cumulatively to a single-tower baseline that consumes multi-view tokens by concatenation, without visual grounding or a reconstruction branch (Table 4). All variants share the same 200K training subset, schedule, and held-out 1K evaluation set.

Table 2: Final shape quality when only the coarse stage is replaced by Coarse3D (+ **Ours**). CD and F-Score measure geometric accuracy, while ULIP and Uni3D measure semantic similarity to the input images. Bold marks the better result of each pair. Note that Pixal3D-MV* uses GT cameras.

Method	**GSO**				**Toys4k**			
	CD ↓	F-Score ↑	ULIP ↑	Uni3D ↑	CD ↓	F-Score ↑	ULIP ↑	Uni3D ↑
TRELLIS-MV	0.0772	0.4021	0.2112	0.2557	0.0397	0.6157	0.2373	0.3610
+ **Ours**	**0.0128**	**0.8413**	**0.2255**	**0.2688**	**0.0131**	**0.8600**	**0.2461**	**0.3690**
TRELLIS2-MV	0.0613	0.5425	0.2073	0.2607	0.0360	0.6522	0.2373	0.3652
+ **Ours**	**0.0127**	**0.8576**	**0.2183**	**0.2650**	**0.0099**	**0.9128**	**0.2391**	**0.3684**
ReconViaGen	0.0785	0.3835	0.2245	0.2707	0.0490	0.5096	0.2519	0.3687
+ **Ours**	**0.0132**	**0.8264**	**0.2391**	**0.2836**	**0.0130**	**0.8537**	**0.2670**	**0.3745**
Pixal3D-MV*	0.0170	0.8003	0.1903	0.2549	0.0127	0.8721	0.2170	0.3495

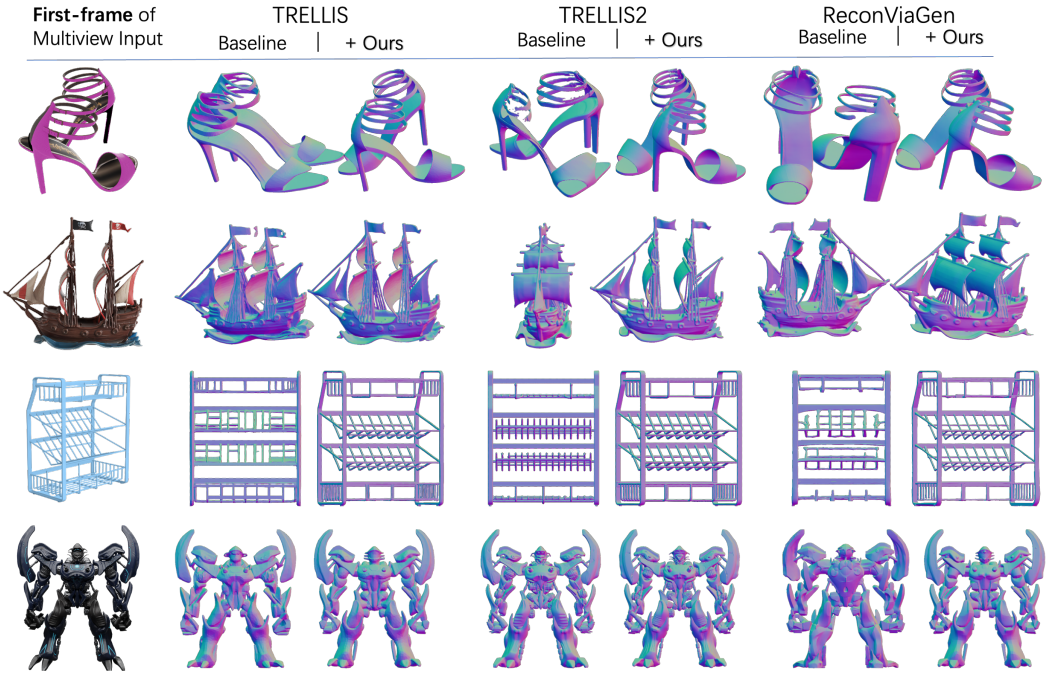


Figure 4: Qualitative comparison across three baselines (TRELLIS, TRELLIS2, and ReconViaGen). For each baseline, we show the original result (*Baseline*) and the result after replacing its first stage with our Coarse3D (+ *Ours*). Anchored by the first frame of the multi-view input (leftmost), our method yields (i) more **controllable orientations**, (ii) higher **structural fidelity**, and (iii) seamless **plug-and-play** integration without modifying subsequent stages.

Table 3: Auxiliary reconstruction accuracy. RRE and RTE are relative camera rotation and translation errors, and CD is measured on the predicted point maps after rigid alignment. * numbers are copied from the original paper. Best in bold.

Method	**GSO**			**Toys4k**		
	RRE ↓	RTE ↓	CD ↓	RRE ↓	RTE ↓	CD ↓
VGGT	10.54	0.2010	0.0226	16.11	0.2352	0.0160
ReconViaGen	4.09	0.0934	0.0203	8.15	0.1293	0.0144
RecGen3D*	2.16	0.0338	–	3.67	0.0637	–
Ours	**2.02**	**0.0319**	**0.0174**	**3.43**	**0.0370**	**0.0128**

Dual tower. Injecting image-tower features layer by layer mainly sharpens structure: CD drops by 58% ($0.0280 \rightarrow 0.0117$) and PA-IoU rises by 0.065, while the IoU–PA-IoU gap remains unchanged ($0.064 \rightarrow 0.068$), showing that grounding alone leaves the global orientation undetermined.

First-frame anchoring. Anchoring collapses the gap by $5.7\times$ to 0.012, with asymmetric gains (IoU $+0.093$ against PA-IoU $+0.038$) that come predominantly from orientation. Because generation and reconstruction now share a single coordinate frame, the auxiliary task also benefits (RRE $5.88 \rightarrow 3.42$, RTE $0.0692 \rightarrow 0.0413$).

Spatial-prior guided attention. With orientation resolved, the remaining gain is structural: PA-IoU rises by 0.085 (the largest gain among the three components) and CD drops by 16%. The attention maps in Fig. 5 show the mechanism: for a representative voxel query, the base model (middle) attends diffusely over the whole object, whereas the guided model (bottom) concentrates on the pixels onto which the query voxel projects, closely following the spatial prior (top). The three designs therefore act on complementary factors: grounding on structure, anchoring on orientation, and guided attention on pixel-level correspondence.

Table 4: Cumulative ablation on the three key components of Coarse3D. Each row adds one component to the row above. The baseline has no image tower and therefore no reconstruction metrics.

Setting	**Coarse Generation**			**Reconstruction**	
	IoU ↑	PA-IoU ↑	CD ↓	RRE ↓	RTE ↓
Baseline	0.4041	0.4677	0.0280	–	–
+ Dual-DiT under canonical space	0.4652	0.5329	0.0117	5.88	0.0692
+ First-frame anchored canonical space	0.5586	0.5704	0.0104	3.42	0.0413
+ Spatial-Prior guided Attention	**0.6392**	**0.6550**	**0.0087**	**3.10**	**0.0399**

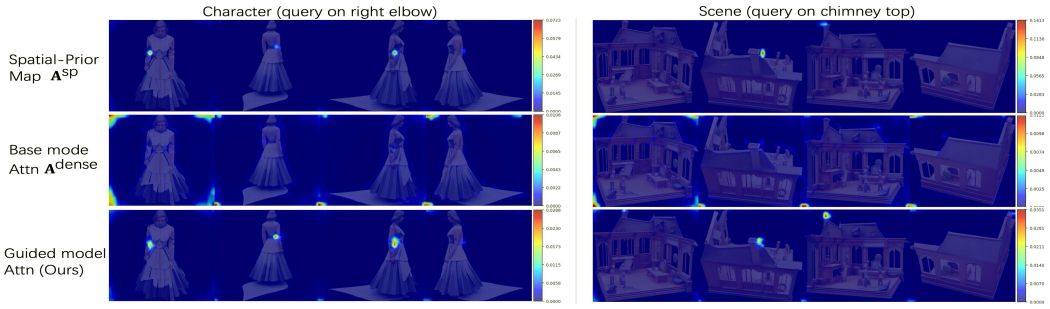


Figure 5: **Effect of spatial-prior guided attention.** For a representative voxel query, we visualize the attention map over pixel tokens of four input views. *Top*: spatial prior $\mathbf{A}^{sp}$ derived from geometric projection. *Middle*: base model attention $\mathbf{A}^{dense}$, which spreads diffusely across the object. *Bottom*: our guided model concentrates on the pixels the query voxel actually projects to.

5 Conclusion

We presented **Coarse3D**, a coarse-stage model for 3D-native generation that targets two limitations of current pipelines: orientation that is uncontrollable because canonicalization is semantically ambiguous, and structures that are inaccurate because 2D-3D grounding is weak. Coarse3D anchors the canonical frame to the first input view, grounds its conditioning features with a dual-tower DiT trained jointly on reconstruction, and regularizes cross-attention with pixel-to-voxel correspondence. Since it preserves the sparse-voxel interface of existing coarse stages, it can replace them without retraining the refinement stage, raising F-Score from 0.54 to 0.86 on GSO and improving final geometry for all three pipelines we tested. Our study focuses on the coarse stage; extending our core designs, e.g., the reconstruction-grounded Dual-DiT and spatial-prior guided attention, to fine-shape generation as a unified framework, though architecturally natural, is left to future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with summarizing and analyzing existing literature and identifying relevant related work, particularly influential papers on image-to-3D generation. For clarity and readability, we also used these tools to assist with translation, English grammar correction, and manuscript polishing. We did not use generative AI tools for any other tasks requiring disclosure, including generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims or contributing to proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning or reformatting datasets, conducting qualitative or thematic analyses. We have reviewed all AI-assisted work to verify its accuracy and preserve the intended meaning. We take full responsibility for the final content of this paper, including all AI-assisted text.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=z0QLeooEEf.

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 (CVPR)*, pp. 7220–7232, June 2026.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.

Laura Downs, Anthony Francis, Nate Maggio, Brandon Cavalcanti, Gerard Tagliabue, Jake Varley, and Brian Ichter. 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.

Tiancheng Fang, Bowen Pan, Lingxi Chen, Jiangjing Lyu, Chengfei Lv, Chaoyue Niu, and Fan Wu. Viaformer: Voxel-image alignment transformer for high-fidelity voxel refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29060–29070, June 2026.

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 (ICCV)*, pp. 14822–14833, October 2025.

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.

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

Zeqiang Lai, Yunfei Zhao, Zibo Zhao, Haolin Liu, Qingxiang Lin, Jingwei Huang, Chunchao Guo, and Xiangyu Yue. Lattice: Democratize high-fidelity 3d generation at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19982–19992, June 2026.

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, SIGGRAPH Conference Papers '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725548. doi: 10.1145/3799902.3811175. URL https://doi.org/10.1145/3799902.3811175.

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 diffusion and interactive geometry refiner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5307–5317, June 2025a.

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*, 2025b.

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

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*, 2025d.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Yichong Lu, Yuzhuo Tian, Zijin Jiang, Yikun Zhao, Yuanbo Yang, Hao Ouyang, Haoji Hu, Huimin Yu, Yujun Shen, and Yiyi Liao. Orientation matters: Making 3d generative models orientation-aligned. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=YVOCwQF9k2.

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.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FjNys5c7VyY.

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*, 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. *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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1798–1808, 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, 2026. URL https://arxiv.org/abs/2605.21121.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=UyNXMqnN3c.

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

S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

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

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

Zehan Wang, Ziang Zhang, Jiayang Xu, Jialei Wang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything v2: Unifying orientation and rotation understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Philip Torr, Xun Cao, and Yao Yao. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *Advances in Neural Information Processing Systems*, 2025.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 21469–21480, June 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 (CVPR)*, pp. 14419–14429, June 2026.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. 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.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25050–25061, October 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*, 42(4):1–16, 2023.

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.

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

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 *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=mrb4eOur7r.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wcaE4Dfgt8.

A Appendix

A.1 Detailed Comparison of 3D Generation Spaces

In the standard preprocessing pipeline for 3D assets, objects are typically centered at the origin and scaled within a unit cube. Subsequently, images are rendered from multiple viewpoints surrounding the object to construct image-3D training pairs. We refer to this classical coordinate system as the Standard Canonical Space (Fig. 2 left), which is defined independently of any viewpoint. Ideally, the object orientation within this space should align with human intuition: the $+z$ axis points upward (opposite to gravity), the $-y$ axis indicates the semantic front of the object, and the object rests naturally on the "ground" formed by the xOy plane. However, methods such as TRELLIS randomly sample rendering images from arbitrary viewpoints during training while keeping the generated target completely unchanged. This forces the generative model to implicitly learn and converge to a unified canonical orientation for each object category. Due to heterogeneous standards across internet-scale datasets, actual object orientations are often highly inconsistent. Specifically, while the gravity direction (determined by both elevation and roll angles) is mostly accurate or can be easily corrected, the azimuth angle (determined by the object's front-facing direction) varies significantly due to differing interpretations among 3D creators. This leads to severe **azimuth ambiguity** within the Standard Canonical Space (Fig. 2 left). OrientationMatters (Lu et al., 2025) highlighted that such inconsistencies severely impair generative training and proposed globally correcting dataset orientations. However, this process is prohibitively labor-intensive. Furthermore, OrientAnythingV2 (Wang et al., 2025b) pointed out that real-world objects may have multiple semantic fronts, making absolute orientation correction physically ill-defined.

To address this issue, Pixal3D (Li et al., 2026) introduced a Pixel-aligned Camera Space (Fig. 2 middle). During training, given a randomly sampled input image with viewing angles $(\theta_a, \theta_e, \theta_r)$ (representing azimuth, elevation and roll), the method applies the identical rotation to the original normalized 3D assets as the final training ground truth. Consequently, the generation result of Pixal3D is strictly aligned with the pose of the input image, which is called "Pixel-aligned 3D Generation". This definition forces the model to actively learn the object's actual pose rather than fitting an ambiguous canonical orientation, thereby improving training stability. Although the Pixel-aligned Camera Space eliminates orientation ambiguity, its solution space covers all possible 3D rotations (i.e., $\mathrm{SO}(3)$ space), substantially increasing the learning difficulty. More critically, introducing elevation and roll transformations **loses the gravity direction** of the generated 3D assets, making them unpredictable for downstream applications such as placing objects plausibly on a ground plane.

Given that azimuth ambiguity is the predominant issue in real-world datasets, we integrate the merits of both above spaces and propose the first-frame anchored canonical space (Fig. 2 right). Specifically, for the first frame with viewing angles $(\theta_a, \theta_e, \theta_r)$ from multi-view inputs, we apply only the azimuth rotation θ_a to the original normalized 3D asset, yielding the final training ground truth. As a result, the azimuth of the training ground truth matches the actual azimuth of the object in the first frame, while the elevation and roll angles maintain the original anti-gravity direction. If the generated azimuth is undesirable, users can rotate it alone without disturbing the gravity direction.

The detailed comparison of the three spaces mentioned above is summarized in Table 5. We argue that our proposed space definition is optimal for three reasons:

- Training robustness: It bypasses the strong assumption of a unified canonical orientation in the Standard Canonical Space, ensuring stable training even on large-scale, inconsistently annotated open-source datasets.
- Reduced learning complexity: Aligning solely along the azimuth dimension significantly shrinks the full $\mathrm{SO}(3)$ space compared to full-view alignment in the Pixel-aligned Camera Space, thereby facilitating the subsequent learning process and improving convergence stability.
- Controllability and intuitive usability: By selecting a desired first image, users can reliably control the azimuth orientation of the generated object. Meanwhile, the elevation and roll angles, which determine the anti-gravity direction, remain consistent with human intuition and world knowledge, as well as standard 3D asset creation conventions.

Table 5: Summary of three types of 3D generative spaces.

Space Type	View-dependent	Alignment of Orientation Angles
Standard Canonical Space	✗	none
Pixel-aligned Camera Space	✓	all (azimuth, elevation and roll)
First-frame Anchored Canonical Space (ours)	✓	only azimuth

A.2 Model and Training Details

A.2.1 Architecture and Initialization

The sparse-structure stage of Coarse3D is a single **Dual-DiT** of about 2.0B parameters, consisting of a Geometry tower and an Image tower that are coupled layer by layer and share the same depth of 30 layers.

Geometry tower. The Geometry tower follows the SS-DiT structure of TRELLIS with 30 layers. It takes voxel latents at 16^3 resolution as input, applies 3D RoPE positional encoding, and flattens them into a sequence of 4096 tokens. Image information is injected through cross-attention at every block, and the tower predicts the Partial View-Aligned coarse structure via flow matching.

Image tower. The Image tower adopts the VGGT architecture but, since the original VGGT has only 24 layers, we extend it to 30 layers to match the depth of the Geometry tower for layer-by-layer feature injection. Each input image is cropped and resized to 512 and encoded by DINOv3, producing 1029 tokens per view. To handle a variable number of views, we introduce view-aware learnable embeddings of shape $[6, 1024]$, supporting up to 6 input views, where 1024 is the DINOv3 image-feature dimension. The tower uses alternating frame-wise and global self-attention, and after each global self-attention layer, its multi-view stereo features are injected into the corresponding layer of the Geometry tower.

Initialization. The Geometry tower is initialized from the pretrained TRELLIS2 SS-DiT which is originally trained with a single image as the condition, and the Image tower from VGGT (Wang et al., 2025a). Weights absent from these checkpoints (e.g., the extended VGGT blocks) are randomly initialized, and the view-aware embedding is initialized to zero.

A.2.2 Training

Stage 1: Base Model Pretraining. We first train the Dual-DiT jointly on the generation and auxiliary reconstruction objectives, with all geometric targets expressed in the first-frame anchored canonical space and $1 \sim 6$ unposed views randomly sampled per iteration. Because the randomly initialized Image tower could otherwise corrupt the pretrained Geometry tower weights, we adopt a warmup strategy over the first 10K steps, during which the weight of the auxiliary reconstruction loss is linearly increased from 0 to 1. After 200K steps at a learning rate of 1×10^{-4}, we obtain the base model.

Stage 2: Attention-Regularized Fine-tuning. We then enable the Spatial-Prior Guided Attention and add its regularization loss, fine-tuning for a further 100K steps at a reduced learning rate of 1×10^{-5} with $\delta = 0.1$, $\beta = 0.2$, and $\lambda_{\text{attn}} = 0.1$. This attention regularizer is applied only during training and introduces no additional cost or camera dependence at inference. All experiments use the AdamW optimizer with a global batch size of 256 on 64 H20 GPUs.

A.3 Evaluation Metrics

We detail the metrics used in the main paper. All CD-based metrics are computed after Umeyama alignment (Umeyama, 1991) to remove residual pose and scale ambiguity.

Table 6: **Ablation on the number of input views.** We vary the number of input images from 1 to 6 and report the coarse structure reconstruction quality of Coarse3D on GSO and Toys4K.

	GSO			Toys4K		
# Views	IoU ↑	PA-IoU ↑	CD ↓	IoU ↑	PA-IoU ↑	CD ↓
1	0.3129	0.3495	0.0285	0.3876	0.4148	0.0236
2	0.4200	0.4534	0.0187	0.5234	0.5375	0.0163
3	0.4834	0.5075	0.0168	0.6045	0.6149	0.0131
4	0.5101	0.5364	0.0124	0.6315	0.6368	0.0091
5	0.5244	0.5449	0.0115	0.6419	0.6465	0.0087
6	0.5320	0.5586	0.0110	0.6484	0.6522	0.0085

Chamfer Distance (CD). We adopt the L1 formulation, which averages nearest-neighbor Euclidean distances in both directions:

$$\mathrm{CD}(\mathcal{P},\mathcal{G}) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\min_{g\in\mathcal{G}}\|p-g\|_2 + \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\min_{p\in\mathcal{P}}\|g-p\|_2. \tag{6}$$

Lower CD indicates better geometric accuracy.

F-Score. F-Score at threshold $\tau = 1\%$ of the ground-truth scale is the harmonic mean of precision (fraction of predicted points within τ of the ground truth) and recall (fraction of ground-truth points recovered within τ). Higher is better.

IoU. Voxel-level intersection over union at resolution 64^3 in the shared canonical space: $\mathrm{IoU} = |V_{\text{pred}} \cap V_{\text{gt}}|/|V_{\text{pred}} \cup V_{\text{gt}}|$.

PA-IoU (PCA-Aligned IoU). Standard IoU is highly sensitive to orientation, penalizing pose ambiguity rather than genuine geometric error. PA-IoU decouples the two by aligning the prediction to the ground truth via principal-axis registration before computing IoU. Concretely, we (i) extract occupied voxels as centered point clouds $X_{\text{pred}}, X_{\text{gt}}$; (ii) obtain their principal frames $U_{\text{src}}, U_{\text{tgt}}$ via SVD of $X^\top X$; (iii) enumerate the $2^3 = 8$ sign flips of U_{src} (inherent to SVD), compute the candidate rotation $R = U_{\text{tgt}}U_{\text{src}}^\top$ for each, and discard mirror candidates with $\det(R) < 0$; (iv) select the R minimizing the bidirectional CD between the rotated prediction and the ground truth; and (v) report the standard IoU under this alignment. The gap between IoU and PA-IoU thus isolates the portion of error attributable to orientation.

Camera Pose (RRE / RTE). Following (Wang et al., 2024), we evaluate camera poses via pairwise Relative Rotation Error and Relative Translation Error, which are invariant to global coordinate ambiguity. For each image pair (i, j):

$$\mathrm{RRE}_{ij} = \arccos\Big(\tfrac{\mathrm{tr}(R_{ij}^\top R_{ij}^\star)-1}{2}\Big), \quad \mathrm{RTE}_{ij} = \arccos\Big(\tfrac{t_{ij}^\top t_{ij}^\star}{\|t_{ij}\|\,\|t_{ij}^\star\|}\Big), \tag{7}$$

where $R_{ij} = R_i R_j^\top$ and t_{ij} denote the predicted relative rotation and translation direction, and $R_{ij}^\star, t_{ij}^\star$ their ground-truth counterparts. Both are reported in degrees, averaged over all image pairs.

Semantic Consistency. We measure semantic similarity between generated shapes and input images using ULIP-2 (Xue et al., 2024) and Uni3D (Zhou et al., 2024) embeddings via cosine similarity. Higher is better.

A.4 Ablation study with more input images

Coarse3D accepts an arbitrary number of input images, which raises a natural question: how does reconstruction quality scale with the number of views? To answer this, we conduct an ablation on Toys4K and GSO by varying the number of input views, and report the results in Tab. 6. Reconstruction quality improves steadily as more views are added, yet the marginal gain shrinks with each additional view, revealing a clear saturation trend once the view count grows large.

A.5 Plug-and-Play Integration with Existing Pipelines

Direct reuse of pretrained VAE representations. A distinctive property of Coarse3D is that it directly reuses the pretrained Voxel VAE (Xiang et al., 2025) *without any retraining*, enabled by our choice of coordinate system. Different from alternatives such as Pixal3D (Li et al., 2026) that redefine geometry in the non-canonical frame and therefore require representations retrained for that frame, our output stays in the same object-centric unit cube as the baseline canonical pose and differs from it only by a rotation about the up-axis. Since the training data of these pipelines already contains objects at different azimuths, upright and in-plane-rotated structures fall within the distribution modeled by both their VAEs and their latent refinement stages.

A representation-agnostic interface. Coarse3D replaces only the first stage of a cascade and communicates with the second stage through a single minimal interface: an occupancy grid over a normalized object-centric voxel space. It is therefore compatible with any voxel-based coarse-to-fine pipeline, regardless of how the refinement stage represents fine geometry—for example, hierarchical voxels (Ren et al., 2024), Flexicubes (He et al., 2025), SLAT (Xiang et al., 2025), O-Voxel (Xiang et al., 2026), T-Voxel (Zhao et al., 2026), or VoxSets (Lai et al., 2026)—and regardless of how that stage is trained. What matters is only that the refinement stage consumes a coarse occupancy at the same resolution and in the same normalized object-centric space. This is precisely why Coarse3D cannot be directly plugged into Pixal3D: although both operate on voxels, Pixal3D's refinement stage lives in a pixel-aligned space, so the interface simply does not match. Note also that, although our coarse structure now lies in the anchored rather than the semantic canonical frame, the refinement stage conditions on both the coarse structure and the input views, which remain mutually consistent under a global change of frame. This is why the three pipelines in Sec. 4.2.2 accept our output without any retraining.