ROAR-3D: ROuting ARbitrary Views for High-Fidelity 3D Generation

Hanxiao Sun[1,2*,◇], Mingxin Yang[2*], Shuhui Yang[2], Zebin He[1,2], Xintong Han[2], Hongbo Fu[1], Chunchao Guo[2†], and Wenhan Luo[1†]

[1] The Hong Kong University of Science and Technology
[2] Tencent Hunyuan

Fig. 1: High-quality 3D geometry generated by **ROAR-3D** from arbitrary collections of concept art, design sketches, and video frames, exhibiting fine geometric detail.

Abstract. Single-image-to-3D generative models can now produce high-quality geometry, yet conditioning on a single view inevitably introduces ambiguity about unseen regions. Multi-view conditioning can reduce this ambiguity, but existing methods either require fixed canonical viewpoints or rely on external reconstruction modules that impose heavy training costs and limit generation quality. We observe that pretrained single-view models already possess strong 2D-to-3D grounding that can be reused for multi-view conditioning. However, a closer analysis reveals that their conditioning mechanism entangles orientation control with geometry transfer, two functions that conflict when images from different

◇ Work done during internship at Tencent Hunyuan.
* Equal contribution.
† Corresponding author.

viewpoints are naively combined. Based on this analysis, we propose **ROAR-3D**, a lightweight method that upgrades a pretrained single-view model to accept an arbitrary number of unposed images. A *token-wise view router* assigns each 3D latent token to its most relevant view, implicitly establishing 2D-to-3D correspondences without explicit pose input. A *dual-stream attention* design preserves the pretrained primary-view behavior while routing auxiliary views through a separate path dedicated to geometric enrichment. An *orientation perturbation strategy* ensures the auxiliary path learns orientation-independent geometry transfer. These components introduce lightweight architectural additions and incur only low additional inference overhead relative to the single-view baseline. ROAR-3D achieves state-of-the-art multi-view 3D generation quality and supports test-time view scaling from 1 to 12+ views with consistent improvements.

Keywords: 3D Generation · Multi-View Conditioning · Diffusion Transformer · View Routing · Pose-Free 3D Reconstruction

1 Introduction

Recent advances in diffusion-based 3D generative models have enabled remarkably high-quality 3D geometry generation from a single image [5, 25, 26, 31, 32, 70, 71, 81, 86, 87], producing increasingly detailed and plausible 3D geometry. By leveraging generative priors learned from large-scale 3D datasets, these methods can synthesize complete geometry even for regions not visible in the input. However, conditioning on a single view inevitably introduces ambiguity: the model must hallucinate unseen structures, and the generated output may not faithfully reflect the user's intended geometry from other viewpoints. This limits the practical utility of single-view methods in applications where the generated 3D geometry must closely match the appearance specified by multiple reference images.

A natural solution is to condition the generation on multiple views. Multi-view conditioned generation occupies a unique position between pure generation and multi-view reconstruction: it retains the ability to complete unobserved regions and tolerate inconsistencies across input views (as it samples from a learned conditional distribution), unlike reconstruction methods [23, 42, 60] that struggle with occlusion, incomplete coverage, weak texture and require strictly consistent multi-view inputs. However, these methods significantly reduce ambiguity by constraining the output to be consistent with multiple observed images. This generative nature also opens up workflows that are inaccessible to traditional reconstruction, such as producing 3D geometry from concept art, design sketches, or video frames (see Fig. 1), substantially reducing the manual effort required in 3D content creation workflows.

Despite their potential, existing multi-view conditioned 3D generation methods face notable limitations, which can be broadly categorized into two categories. The first category adopts fixed viewpoint configurations, requiring input

images to be captured from a predefined set of canonical poses. This includes feed-forward reconstruction-generation hybrids [16, 52, 73] as well as commercial systems [14, 57, 58] that only support a small number of predetermined viewpoints, severely limiting their applicability to unconstrained real-world inputs. The second attempts to incorporate flexible multi-view information by integrating external reconstruction modules such as VGGT [60] features into the generative pipeline [1]. However, the domain gap between reconstruction features and the latent generative space imposes a substantial training overhead. Furthermore, the generation fidelity is strictly upper-bounded by the accuracy of the reconstruction module; consequently, when external features are inaccurate or incomplete, the generative model lacks the inherent robustness to rectify these artifacts or plausibly complete the missing regions. A truly flexible system should accept an arbitrary number of images from arbitrary viewpoints without requiring camera poses or external reconstruction as input.

We observe that pretrained single-view image-to-3D models already contain the necessary knowledge for multi-view conditioning. Given an input image captured from a wide range of viewpoints, these models can map it to plausible canonical 3D geometry, suggesting that the attention layers bridging 2D image tokens and 3D latent tokens have learned useful 2D–3D correspondences across viewing angles. These layers are among the most expensive components to train, as they must establish correspondences across two very different modalities. Rather than re-learning these correspondences from scratch, we aim to maximally reuse them. However, naively feeding multiple views into the existing attention layers leads to degraded results. A closer analysis reveals that the conditioning process entangles two distinct roles: (i) defining the global orientation of the output, and (ii) transferring local geometric structure. In the single-view case, this coupling is harmless, but with multiple views, conflicting orientation signals from different viewpoints destabilize the generation. The challenge, therefore, reduces to: *how to let multiple views contribute complementary geometry without disrupting the learned weights or introducing orientation conflicts?*

We introduce **ROAR-3D**, a lightweight method that upgrades a pretrained single-image-to-3D model to accept arbitrary multi-view input without training the backbone from scratch or introducing external encoders. At its core is a *token-wise view router* that, for each 3D latent token, selects the single most informative view before the attention computation. Because each 3D latent token carries spatial meaning, this selection naturally establishes 2D-to-3D correspondences across views, sidestepping the need for explicit pose conditioning. The router's projections are initialized from the pretrained attention weights, greatly reducing training cost. To separate orientation control from geometric structure transfer, we adopt a *dual-stream attention* design: the *primary stream* preserves the pretrained behavior for the reference view (setting both orientation and base geometry), while the *auxiliary stream* handles supplementary views (enhancing geometric structure without altering orientation). An *orientation perturbation strategy* further strengthens the auxiliary stream by training it on deliberately misaligned 3D latents, ensuring it learns orientation-independent ge-

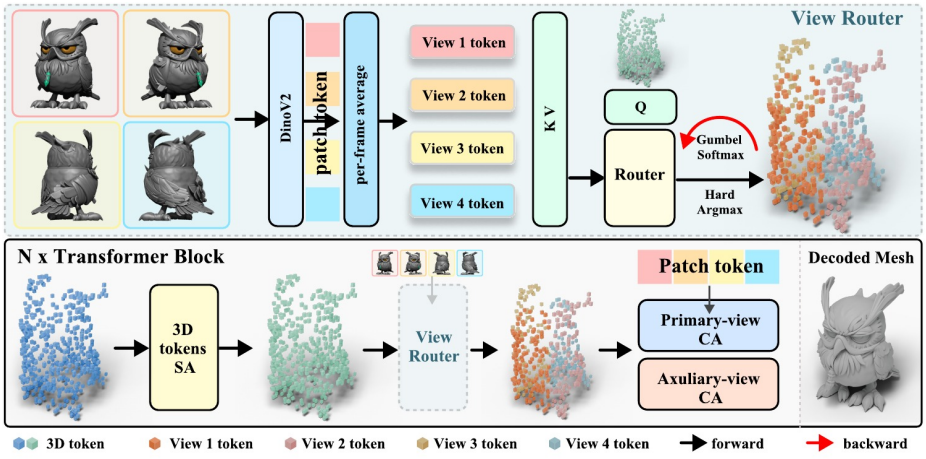

Fig. 2: An overview illustration of the proposed **ROAR-3D** framework, which seamlessly integrates supplementary visual cues from arbitrary unposed views with pretrained single-view generative priors via a lightweight token-wise routing mechanism for high-fidelity 3D reconstruction.

ometric transfer. Together, these components keep the per-token cross-attention cost identical to the single-view baseline and incur only low additional overhead from view encoding and routing, while enabling the model to dynamically fuse information from any number of unposed views.

Our contributions are as follows:

- We propose **ROAR-3D**, a method that extends pretrained single-view 3D generative models to support multi-view input without requiring external encoders or camera poses.
- We introduce a token-wise view router that uses 3D latent tokens as anchors to ground multi-view 2D information in 3D space, combined with a dual-stream attention design and orientation perturbation strategy that explicitly decouples orientation control from geometric structure transfer.
- ROAR-3D achieves state-of-the-art multi-view 3D generation quality with low additional inference overhead, and supports test-time view scaling from 1 to 12+ views with consistent improvements.

2 Related Work

Single-View 3D Generation. Single-view 3D generation aims to lift a single 2D image into a complete 3D representation, often via diffusion models [15, 24, 35, 48, 82, 83]. Early approaches distill 3D knowledge from pretrained 2D diffusion models via score distillation sampling [33, 46, 47, 53, 54, 63], while another line of work generates multi-view images from a single input and fuses them into 3D representations [27, 28, 36, 37, 52, 64, 66, 68, 73, 75, 76, 91]. To achieve high-fidelity

appearance, several approaches specifically focus on texture generation and material decomposition, leveraging generative priors to map 2D appearance onto 3D manifolds [2,9,13,38,40]. More recently, 3D native generative methods apply diffusion processes directly on 3D representations such as point clouds [41, 44, 90], voxel grids [18, 43, 55], triplanes [3, 50, 62], and 3D Gaussians [85]. Latent 3D diffusion models [31, 32, 69, 81, 84, 86–88] further improve quality by learning mappings between 2D images and compressed 3D latent spaces. These models have established strong 2D-to-3D grounding through large-scale training; however, they are inherently limited by single-view ambiguity and must hallucinate all unseen geometry. Our work builds directly on this family of models, aiming to activate their learned grounding capability for multi-view conditioning.

Multi-View 3D Reconstruction. Multi-view reconstruction recovers 3D geometry from multiple calibrated or uncalibrated images. Classical multi-view stereo (MVS) methods triangulate correspondences across views [10, 11, 49, 74], while learning-based approaches [4, 6, 12, 29, 30, 59, 78–80] improve efficiency via deep networks. NeRF-based methods [34, 42, 65] and 3D Gaussian Splatting [17, 23, 51] enable high-quality novel view synthesis but typically require accurate camera poses. Recent works such as DUSt3R [61] and VGGT [60] estimate point clouds and poses directly from unposed views. Large Reconstruction Models (LRMs) [16,52,66,73,75] regress compact 3D representations from multi-view inputs in a feed-forward manner. Despite these advances, reconstruction methods fundamentally require consistent multi-view observations and struggle with occlusion, weak texture, and incomplete coverage. Without generative priors, they tend to produce incomplete geometry or over-smoothed surfaces in unobserved regions.

Multi-View Conditioned 3D Generation. Multi-view conditioned 3D generation combines the completeness of generative models with the accuracy afforded by multiple observations. Existing methods fall into two categories. The first adopts fixed viewpoint configurations: feed-forward models [16, 52, 73] and commercial systems [14, 20, 21, 58] accept images only from a small set of predetermined canonical poses, limiting their applicability to unconstrained inputs. The second category integrates external reconstruction modules into the generative pipeline. For instance, [1] injects VGGT [60] features into a diffusion model to provide reconstruction priors. While effective, the domain gap between reconstruction features and the latent generative space imposes a heavy training burden, and generation quality is bounded by the accuracy of the external module. Other works explore optimization-based pose alignment [67] or volume diffusion [36, 72], but face trade-offs between computational cost and output quality. In contrast, ROAR-3D requires no external reconstruction module, no camera poses, and no fixed viewpoint assumption. By introducing a lightweight routing mechanism and dual-stream attention into the pretrained backbone, it enables direct and efficient conditioning on arbitrary unposed multi-view images.

3 Method

We present **ROAR-3D**, a lightweight and effective method that extends a pre-trained single-image-to-3D generative model to accept an arbitrary number of multi-view images as input, so that the generated 3D geometry faithfully captures the structural details specified by every input view. Illustrated in Fig. 2, we build upon the flow-matching Diffusion Transformer (DiT) and the 3DShape2VecSet-based VAE of Hunyuan3D 2.1 [19]. To incorporate multi-view information without retraining the backbone from scratch, we introduce three key components: a **token-wise view routing mechanism** (§3.1) that assigns each 3D latent token to its most informative view by repurposing the pre-trained cross-attention module; a **dual-stream cross-attention design** (§3.2) where the primary branch uses a reference view to set 3D orientation and establish its corresponding geometric structure, while the auxiliary branch adds geometric details from other views without changing geometry orientation; and an **orientation perturbation strategy** (§3.3) that forces the auxiliary branch to learn an orientation-invariant image-to-3D geometric transfer, while keeping orientation control strictly in the primary branch. We further adopt a second-stage refinement model based on LATTICE [26] with the same multi-view conditioning architecture to enhance the quality of the generated geometry (§3.4).

3.1 Token-Wise View Router

In modern image-to-3D DiT models [5, 19, 26, 70, 71, 87], the attention layers that bridge 2D image tokens and 3D latent tokens must learn correspondences across different modalities, making them one of the most challenging components to train. Since a well-trained single-view model can map images captured from diverse viewpoints to plausible canonical 3D geometry, we argue that it has learned useful 2D–3D correspondences across viewing angles. The challenge for multi-view extension is therefore not about re-learning these correspondences from scratch, but about designing a conditioning mechanism that lets multiple views contribute without disrupting the learned weights.

A naïve approach of concatenating all view tokens and feeding them into a single attention layer introduces two problems: (i) each 3D token must attend to tokens from all views simultaneously, creating ambiguity from conflicting visual signals across views, and (ii) computational cost grows linearly with the number of views. Alternatively, encoding view identity, whether through fixed viewpoint bins with learnable embeddings or through estimated camera parameters, either restricts the model to a predefined set of poses or introduces errors from inaccurate pose estimation. Our router takes a different path: each 3D latent token directly selects its most informative view based on feature affinity, effectively using the 3D tokens themselves as anchors that ground multi-view 2D information in 3D space.

To be specific, the router is applied independently at each transformer block. For each 3D latent token, it computes a routing score against every input view and selects the top-scoring one *before* cross-attention, keeping the per-token

attention cost identical to the single-view baseline. Since the router's task of determining which view best matches a given 3D token is closely related to what the pretrained cross-attention already computes at the token level, we initialize the router's query and key projections directly from the pretrained cross-attention weights, which greatly reduces the training cost of the router. We now describe its architecture in detail.

Router architecture. As illustrated in Fig. 2, given a set of V input views whose DINOv2 [45] patch features are $\{\mathbf{F}_v \in \mathbb{R}^{S\times D}\}_{v=1}^{V}$ and 3D latent tokens $\mathbf{Z} \in \mathbb{R}^{N\times D}$, the router at each transformer block ℓ maintains query and key projection matrices $\mathbf{W}_q^{(\ell)}, \mathbf{W}_k^{(\ell)}$ initialized as described above, along with the original QK-normalization (RMSNorm as in [19]). The router query is derived from each 3D latent token. For the router key, since the router selects among views rather than individual patches, we need a single view-level representation for each input view. A natural candidate is the CLS token, but it encodes a global image summary that does not capture how individual patches relate to 3D tokens. Instead, we use the mean-pooled DINOv2 patch tokens: views with more patches correlated to a given 3D token produce a mean representation naturally closer to it in feature space, yielding more accurate routing: Let $\bar{\mathbf{f}}_v = \frac{1}{S}\sum_{s=1}^{S} \mathbf{f}_{v,s}$ denote the mean-pooled patch representation of view v, and $\tilde{\mathbf{z}}_i^{(\ell)} = \mathrm{LN}(\mathbf{z}_i^{(\ell)})$ the layer-normalized 3D latent token. The router query and key are:

$$\mathbf{q}_i^{(\ell)} = \mathrm{RMSNorm}\Big(\mathbf{W}_q^{(\ell)}\,\tilde{\mathbf{z}}_i^{(\ell)}\Big), \quad \mathbf{k}_v^{(\ell)} = \mathrm{RMSNorm}\Big(\mathbf{W}_k^{(\ell)}\,\bar{\mathbf{f}}_v\Big), \tag{1}$$

where $\mathbf{W}_q^{(\ell)} \in \mathbb{R}^{Hd\times D}$ and $\mathbf{W}_k^{(\ell)} \in \mathbb{R}^{Hd\times D}$ project the inputs into a lower-dimensional space shared with the pretrained cross-attention. Both the pre-projection LayerNorm and post-projection RMSNorm are inherited from the pretrained cross-attention of block ℓ.

Routing logits. Both $\mathbf{q}_i^{(\ell)}$ and $\mathbf{k}_v^{(\ell)}$ are split into H heads of dimension d, yielding $\mathbf{q}_{i,h}^{(\ell)}, \mathbf{k}_{v,h}^{(\ell)} \in \mathbb{R}^d$. The routing logit is:

$$r_{i,v}^{(\ell)} = \mathbf{w}_{\mathrm{agg}}^{\top}\left[\frac{(\mathbf{q}_{i,1}^{(\ell)})^{\top}\mathbf{k}_{v,1}^{(\ell)}}{\sqrt{d}},\ \ldots,\ \frac{(\mathbf{q}_{i,H}^{(\ell)})^{\top}\mathbf{k}_{v,H}^{(\ell)}}{\sqrt{d}}\right], \tag{2}$$

where $\mathbf{w}_{\mathrm{agg}} \in \mathbb{R}^H$ is learnable, initialized to $1/H$.

Discrete selection via Gumbel-Softmax. Each token must select exactly one view, but arg max is not differentiable. We use hard Gumbel-Softmax [22] with straight-through estimation. During training, Gumbel noise is added to encourage exploration:

$$v_i^* = \arg\max_v \Big(r_{i,v}^{(\ell)} + g_v\Big), \quad g_v = -\log(-\log(u_v)),\ u_v \sim \mathrm{Uniform}(0,1). \tag{3}$$

Gradient flow is enabled via:

$$\mathbf{y}_i = \mathbf{y}_i^{\text{hard}} - \text{sg}(\mathbf{y}_i^{\text{soft}}) + \mathbf{y}_i^{\text{soft}}, \tag{4}$$

where $\mathbf{y}_i^{\text{hard}} \in \{0,1\}^V$ is the one-hot selection at v_i^*, $\mathbf{y}_i^{\text{soft}} = \text{softmax}((\mathbf{r}_i^{(\ell)} + \mathbf{g})/\tau)$, and $\text{sg}(\cdot)$ is stop-gradient. The forward pass uses the discrete $\mathbf{y}_i^{\text{hard}}$; gradients in the backward pass flow through $\mathbf{y}_i^{\text{soft}}$. At inference, noise is removed and selection is deterministic.

The selected view v_i^* and soft weights $\mathbf{y}_i$ are then passed to the dual-stream cross-attention (§3.2).

3.2 Dual-Stream Cross-Attention

In single-view image-to-3D generation, the attention layers that bridge 2D image tokens and 3D latent tokens implicitly serve a dual role: they align the output 3D geometry to a canonical orientation determined by the input viewpoint, and they transfer geometric structure from the 2D image into the 3D representation. These two functions are entangled within a single attention module. When extending to multi-view conditioning, however, they must be factored apart: a designated *primary view* should anchor the output orientation and establish the corresponding geometric structure, while the remaining *auxiliary views* should only contribute additional geometric structure without altering the established orientation.

We realize this factorization through a dual-stream architecture. Two structurally identical cross-attention modules are maintained: the *primary stream* $\text{CA}_{\text{p}}^{(\ell)}$, initialized from the pretrained weights, which preserves the original dual role described above; and the *auxiliary stream* $\text{CA}_{\text{a}}^{(\ell)}$, initialized by copying all parameters from $\text{CA}_{\text{p}}^{(\ell)}$, which is restricted to geometric structure transfer only. One input view is designated as the primary view v_p; at inference, this view is specified by the user or taken as the first input view, and it anchors the canonical orientation of the output. The router assigns each 3D latent token to a specific view; tokens assigned to the primary view are processed by $\text{CA}_{\text{p}}^{(\ell)}$, while tokens assigned to any auxiliary view are processed by $\text{CA}_{\text{a}}^{(\ell)}$. Formally, the cross-attention output for token i at layer ℓ is:

$$\mathbf{o}_i^{(\ell)} = \begin{cases} \text{CA}_{\text{p}}^{(\ell)}(\tilde{\mathbf{z}}_i^{(\ell)},\ \mathbf{F}_{v_i^*}) & \text{if } v_i^* = v_p, \\ \text{CA}_{\text{a}}^{(\ell)}(\tilde{\mathbf{z}}_i^{(\ell)},\ \mathbf{F}_{v_i^*}) & \text{otherwise,} \end{cases} \tag{5}$$

where v_i^* is the view selected by the router (§3.1) and $\mathbf{F}_{v_i^*}$ denotes the patch features of the selected view. To close the gradient path back to the router, the output is modulated by the soft weight from the Gumbel-Softmax:

$$\hat{\mathbf{o}}_i^{(\ell)} = \mathbf{o}_i^{(\ell)} \cdot y_{i,v_i^*}, \tag{6}$$

where $y_{i,v_i^*} = 1$ in the forward pass (output unchanged) but carries the soft gradient in the backward pass. When only a single view is provided, all tokens

are routed to $\mathrm{CA}_{\mathrm{p}}^{(\ell)}$ and the model reduces to standard single-view generation. The only newly introduced parameters are $\mathrm{CA}_{\mathrm{a}}^{(\ell)}$ (a copy of the original module) and the lightweight router (§3.1).

3.3 Orientation Perturbation

While CA_{p} closely mirrors the pretrained single-view attention and requires little adaptation, CA_{a} needs to decouple orientation control from geometric structure transfer, which requires additional training effort beyond standard fine-tuning.

A key difficulty is that, in standard training, the geometry latent is always canonically aligned with the primary view. Under this setting, CA_{a} can take a shortcut by relying on this fixed alignment rather than learning orientation-independent geometric transfer, causing it to fail when the 3D orientation and view directions do not follow the canonical relationship at inference.

To address this, we introduce an orientation perturbation strategy that constructs training samples specifically designed to strengthen CA_{a}. With probability p_{pert}, the following three operations are jointly applied during training: **(i) Misaligned 3D latent construction.** The 3D point cloud is randomly rotated to one of $\{0°, 90°, 180°, 270°\}$ in azimuth (with elevation fixed at zero), while explicitly excluding any rotation that would align it with an input view, ensuring the 3D latent is *not* canonically aligned with any conditioning image. **(ii) Primary view removal.** The primary view designation is discarded, and all input views are treated as auxiliary, so that every view is processed exclusively through CA_{a}. **(iii) Selective weight update.** Only CA_{a} is updated in this mode; CA_{p} remains frozen. By jointly applying these operations, CA_{a} is forced to establish 2D–3D correspondence purely through content-level feature affinity, independent of any orientation prior.

3.4 Geometric Refinement Stage

To further improve the quality and cleanness of the generated geometry, we adopt a second-stage refinement model following LATTICE [26]. This refinement stage uses the same multi-view conditioning architecture and training strategy described above, including the token-wise view router, dual-stream attention, and orientation perturbation. These components are applied to a higher-resolution geometric latent space. The coarse geometry produced by the first stage provides positional encoding for the refinement model through a voxel grid, guiding the generation of fine-grained geometric details.

4 Experiments

4.1 Experimental Setup

Dataset. We train and evaluate ROAR-3D on a filtered subset of Objaverse [8] and ObjaverseXL [7], comprising approximately 300K high-quality 3D objects.

We follow the data processing pipeline of Hunyuan3D 2.1 [19], including watertight mesh conversion, point cloud sampling with normals, and conditional image rendering. For rendering, we divide the azimuth range into four bins centered at $\{0°, 90°, 180°, 270°\}$, each spanning $\pm 45°$. Within each bin, we render 16 images with randomly sampled field of view (20°–70°) and elevation (−90°–90°). We hold out 100 objects that are not seen during training as the Objaverse/ObjaverseXL test set for geometry-based evaluation.

To assess generalization to real-world captures, we construct a multi-source evaluation set of 200 objects, ANYVIEW-200. This set comprises 120 objects with multi-view images generated by Nano Banana Pro [56] (which may contain cross-view inconsistencies inherent to AI generation), 70 objects collected from the internet, and 10 objects captured from real-world scenes. The number of input views per object ranges from 1 to 12, and the views exhibit diverse real-world challenges, including strong perspective distortion, varying camera intrinsics across views of the same object, and incomplete coverage of the object surface. Since ANYVIEW-200 has no ground-truth meshes, it is used for qualitative evaluation and semantic image-shape similarity rather than geometry metrics.

Evaluation Metrics. We adopt a comprehensive set of metrics spanning both geometric accuracy and semantic quality. For geometry, we report symmetric Chamfer Distance (CD, $\times 10^{-3}$) and F-Score (F1, %) at thresholds of 0.1 and 0.05 on the held-out Objaverse/ObjaverseXL test set. We sample 100K points from each mesh, normalize meshes by their bounding boxes, and align predictions to ground truth using ICP before computing CD and F1. To assess the semantic quality of generated geometry, we report ULIP [77] and Uni3D [89] image similarities (ULIP-I and Uni-I), which measure the alignment between the generated 3D shape and the input image condition in a shared embedding space. For multi-view inputs, we compute the image-shape similarity against each input view and average the scores over all views.

Baseline Methods. We compare ROAR-3D against the following categories of methods: (1) *Single-view generative models*: Hunyuan3D 2.1 [19], Trellis [71], Trellis2 [70]; (2) *Multi-view reconstruction methods*: LGM [52] and VGGT [60]; (3) *Multi-view conditioned generation*: ReconViaGen [1], a TRELLIS-based extension that injects external VGGT reconstruction features into the generative pipeline. We additionally include qualitative comparisons with two leading commercial 3D generation services (denoted as Model 1 and Model 2 in Fig. 3). These services are anonymized due to their terms of use and are included as illustrative qualitative comparisons rather than controlled benchmark evidence. For single-view baselines, we provide the same reference view used by ROAR-3D. For multi-view methods, we provide all V available views.

Implementation Details.

First Stage (Coarse Geometry). We initialize the backbone 3DShape2VecSet VAE and DiT from the pretrained weights of Hunyuan3D 2.1 [19]. The DiT has a hidden dimension of 2048 and operates on 4096 latent tokens. The image features are extracted from a frozen DINOv2 ViT-L/14 [45] encoder, yielding S=256 patch tokens of dimension D=1024 per view, which are projected to match the DiT hidden dimension. The router projection matrices $\mathbf{W}_q^{(\ell)}, \mathbf{W}_k^{(\ell)}$ and the auxiliary attention path $\mathrm{CA_a}$ are initialized by copying the corresponding pretrained weights. The Gumbel-Softmax temperature τ is set to 1.0 throughout training. During training, we randomly sample 1 to 4 auxiliary views per instance with uniform probability, allowing the model to handle variable view counts. We train with the AdamW optimizer [39] at a learning rate of 1×10^{-5} with a cosine decay schedule. The orientation perturbation probability is set to $p_{\mathrm{pert}} = 0.2$. The model is trained for 50K iterations with a per-GPU batch size of 4, requiring approximately 3∼4 GPU days. In this stage, the DiT, primary stream, auxiliary stream, and router are trainable (3.45B parameters), while the DINOv2 encoder and VAE are frozen (606M parameters). At test time, the model generalizes to an arbitrary number of views without modification, supporting scaling well beyond the training range (see supplementary for more comparison).

Second Stage (Geometric Refinement). We build upon LATTICE [26] with a 1.6B-parameter DiT operating on 6144 latent tokens, replacing its self-attention-based conditioning with cross-attention to enable integration of our multi-view modules (view router, dual-stream attention, and orientation perturbation). The image encoder is upgraded to DINOv2 ViT-G/14 [45]; all other training settings follow LATTICE. This stage contains 4.53B trainable parameters and 1.44B frozen parameters, and requires approximately 10 GPU days.

4.2 Comparison

Quantitative Results. Tab. 1 summarizes results on the held-out test set of 100 objects. Single-view methods (Hunyuan3D 2.0, Trellis2) achieve reasonable quality but suffer from high CD due to hallucinated geometry in unseen regions. Multi-view reconstruction methods reduce this ambiguity yet remain limited: LGM produces over-smoothed outputs with the lowest F1 scores across both thresholds, and VGGT, despite better geometric coverage, scores poorly on semantic metrics (ULIP-I: 0.103) due to the absence of a generative prior. Among multi-view conditioned methods, Hunyuan3D 2.0-MV shows marginal improvement over its single-view counterpart, and ReconViaGen achieves the best baseline performance with the lowest CD (32.055) and highest semantic scores (Uni-I: 0.345), though the geometric gap over single-view methods remains modest. ROAR-3D stage 1 already outperforms all baselines across every metric (CD: 25.372 vs. 32.055 for ReconViaGen), and stage 2 refinement brings further substantial improvement, reducing CD to 21.039 and raising F1(0.05) to 81.6—a 34% CD reduction and 12% F1(0.05) improvement over the strongest baseline.

Table 1: Quantitative comparison on the held-out Objaverse/ObjaverseXL test set of 100 objects. Best results are in **bold**, second best are underlined. CD is multiplied by 10^3 (lower is better). Higher is better for all other metrics.

Method	CD↓	F1(0.1)↑	F1(0.05)↑	ULIP-I↑	Uni-I↑
Single-view generative models					
Hunyuan3D 2.0 [19]	35.421	80.5	64.2	0.141	0.302
Hunyuan3D 2.1 [19]	34.031	82.6	68.7	0.147	0.317
Trellis2 [70]	33.722	84.4	71.0	0.149	0.321
Multi-view reconstruction					
LGM [52]	58.009	71.0	52.3	0.089	0.215
VGGT [60]	34.686	81.3	65.4	0.103	0.211
Multi-view conditioned generation					
Hunyuan3D 2.0-MV [19]	33.954	82.4	67.1	0.147	0.301
ReconViaGen [1]	32.055	85.1	72.7	0.165	0.345
ROAR-3D (Ours) - stage 1	25.372	88.7	76.8	0.168	0.346
ROAR-3D (Ours) - stage 2	**21.039**	**91.2**	**81.6**	**0.173**	**0.349**

Qualitative Results. Fig. 3 presents qualitative comparisons across representative examples. Several trends are visible. First, ROAR-3D produces geometrically correct structures where other methods falter: the hammer (row 2) maintains its correct upright orientation, the aircraft (row 4) faithfully reproduces both the nose and tail geometry, and later rows show that our outputs consistently avoid the distorted or tilted shapes that appear in some baselines. Second, our method achieves higher detail fidelity with the input views: the warrior (row 1) preserves fine-grained surface details that match the input, the portrait (row 6) correctly reconstructs the pendant on the back of the jacket visible in the reference image, and the cartoon characters (rows 7–8) retain ornamental details consistent across views. Third, ROAR-3D demonstrates robustness to inconsistent inputs: the crocodile (row 3) is captured from views with noticeable appearance discrepancies, yet our method produces a physically plausible and complete geometry while maximizing fidelity to each input view. Overall, by combining the generative prior of the pretrained DiT with accurate multi-view conditioning, ROAR-3D yields complete 3D shapes with correct global orientation and faithful local details, outperforming both reconstruction-based and generation-based baselines.

4.3 Ablation Study

We conduct ablation studies on the held-out Objaverse/ObjaverseXL test set to validate the key design choices of ROAR-3D. To isolate the effect of our multi-view conditioning module, all ablations are performed using the first-stage model only (without the lattice refinement stage). Unless otherwise stated, all variants are trained under identical settings for 50K iterations.

Fig. 3: Qualitative comparison with baseline methods. ROAR-3D produces complete, high-fidelity 3D shapes that are consistent with all input views.

Component Ablation. We incrementally add each proposed component to a baseline that simply concatenates all multi-view tokens and feeds them into the original attention layers. The three components are added in order: (1) the *token-wise view router*, which replaces naive concatenation with per-token view selection; (2) *dual-stream attention*, which separates the primary-view path from the auxiliary-view path; and (3) the *orientation perturbation strategy*, which yields our full model. Quantitative results are reported in Tab. 2, and qualitative comparisons are shown in Fig. 4.

As shown in Fig. 4, each added component brings visible improvement in orientation accuracy and geometric fidelity. For the fighter jet (row 1), the baseline incorrectly replicates the nose geometry onto the tail; introducing the router partially alleviates this, and adding dual-stream attention further improves struc-

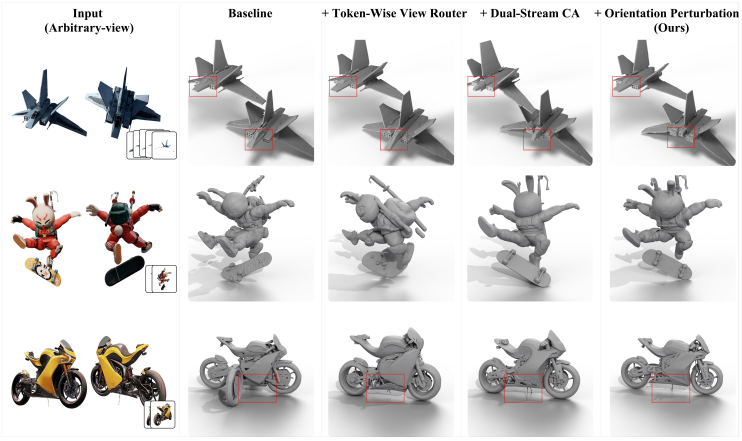

Fig. 4: Ablation study on the three key components of ROAR-3D. Removing any single component leads to incorrect orientation or degraded geometry (highlighted with red boxes). Best viewed when zoomed in.

tural separation, but only the full model with orientation perturbation correctly distinguishes front and rear. For the cartoon character (row 2), the baseline produces both incorrect orientation and degraded details; each component progressively improves orientation control and geometric faithfulness, with the full model being the only variant that achieves both. For the motorcycle (row 3), the baseline generates duplicated wheels; the router resolves this, but the kickstand geometry remains incorrect until dual-stream attention and orientation perturbation are both applied. We note that some intermediate variants still exhibit orientation errors; for fair visual comparison, we manually rotate them to approximately matching viewpoints.

Table 2: Ablation on the three key components of ROAR-3D.

Variant	CD↓	F1(0.1)↑	F1(0.05)↑	ULIP-I↑	Uni-I↑
Baseline	39.412	79.6	66.3	0.154	0.322
+ Token-wise View Router	28.143	86.3	73.5	0.159	0.335
+ Dual-Stream Attention	26.732	87.5	75.1	0.163	0.339
+ Orientation Perturbation (Ours)	**25.372**	**88.7**	**76.8**	**0.168**	**0.346**

5 Conclusion

We presented ROAR-3D, a lightweight method that upgrades pretrained single-view 3D generative models to accept arbitrary unposed multi-view input. Three complementary components, a token-wise view router, dual-stream attention, and orientation perturbation, decouple orientation control from geometry transfer, enabling effective multi-view fusion without external modules or camera poses, while incurring only low additional inference overhead. ROAR-3D achieves state-of-the-art quality and supports test-time view scaling from 1 to 12+ views.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62372480), in part by 2025 Tencent AI Lab Rhino-Bird Focused Research Program, in part by HKUST-MetaX Joint Lab Fund (No. METAX24EG01-D), and in part by a grant from the NSFC/RGC Collaborative Research Scheme sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China and National Natural Science Foundation of China (Project No. CRS_HKUST605/25).

References

1. Chang, J., Ye, C., Wu, Y., Chen, Y., Zhang, Y., Luo, Z., Li, C., Zhi, Y., Han, X.: Reconviagen: Towards accurate multi-view 3d object reconstruction via generation (2025), https://arxiv.org/abs/2510.23306
2. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 18558–18568 (2023)
3. Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L., Su, H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In: ICCV. pp. 2416–2425 (2023)
4. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: ICCV. pp. 1538–1547 (2019)
5. Chen, Y., Li, Z., Wang, Y., Zhang, H., Li, Q., Zhang, C., Lin, G.: Ultra3d: Efficient and high-fidelity 3d generation with part attention. arXiv preprint arXiv:2507.17745 (2025)
6. Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L.E., Ramamoorthi, R., Su, H.: Deep stereo using adaptive thin volume representation with uncertainty awareness. In: CVPR. pp. 2524–2534 (2020)
7. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. NeurIPS **36** (2024)
8. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: ICCV. pp. 13142–13153 (2023)
9. Feng, Y., Yang, M., Yang, S., Zhang, S., Yu, J., Zhao, Z., Liu, Y., Jiang, J., Guo, C.: Romantex: Decoupling 3d-aware rotary positional embedded multi-attention network for texture synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17203–17213 (2025)
10. Furukawa, Y., Hernández, C., et al.: Multi-view stereo: A tutorial. Foundations and trends® in Computer Graphics and Vision **9**(1-2), 1–148 (2015)
11. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: ICCV. pp. 873–881 (2015)
12. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: CVPR. pp. 2495–2504 (2020)
13. He, Z., Yang, M., Yang, S., Tang, Y., Wang, T., Zhang, K., Chen, G., Liu, Y., Jiang, J., Guo, C., et al.: Materialmvp: Illumination-invariant material generation via multi-view pbr diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26294–26305 (2025)

14. Hitem3D Team: Hitem3d: High-quality 3d model generation service (2024), https://www.hitem3d.ai/, accessed: 2024-05-20
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
16. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. In: ICLR (2023)
17. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024)
18. Hui, K.H., Li, R., Hu, J., Fu, C.W.: Neural wavelet-domain diffusion for 3d shape generation. In: SIGGRAPH Aisa. pp. 1–9 (2022)
19. Hunyuan3D, T., Yang, S., Yang, M., Feng, Y., Huang, X., Zhang, S., He, Z., Luo, D., Liu, H., Zhao, Y., Lin, Q., Lai, Z., Yang, X., Shi, H., Zhao, Z., Zhang, B., Yan, H., Wang, L., Liu, S., Zhang, J., Chen, M., Dong, L., Jia, Y., Cai, Y., Yu, J., Tang, Y., Guo, D., Yu, J., Zhang, H., Ye, Z., He, P., Wu, R., Wei, S., Zhang, C., Tan, Y., Sun, Y., Niu, L., Huang, S., Zheng, B., Liu, S., Chen, S., Yuan, X., Yang, X., Liu, K., Zhu, J., Chen, P., Liu, T., Wang, D., Liu, Y., Linus, Jiang, J., Huang, J., Guo, C.: Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material (2025), https://arxiv.org/abs/2506.15442
20. Hy-3D Team: Hy-3d (2024), https://hy-3d.com, accessed: 2024-05-20
21. Hyper3D Team: Hyper3d: High-fidelity 3d asset generation (2024), https://hyper3d.ai/, accessed: 2024-05-20
22. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. TOG **42**(4), 139–1 (2023)
24. Kong, Z., Gao, F., Zhang, Y., Kang, Z., Wei, X., Cai, X., Chen, G., Luo, W.: Let them talk: Audio-driven multi-person conversational video generation. Advances in Neural Information Processing Systems **38**, 70990–71013 (2026)
25. Lai, Z., Zhao, Y., Liu, H., Zhao, Z., Lin, Q., Shi, H., Yang, X., Yang, M., Yang, S., Feng, Y., et al.: Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. arXiv preprint arXiv:2506.16504 (2025)
26. Lai, Z., Zhao, Y., Zhao, Z., Liu, H., Lin, Q., Huang, J., Guo, C., Yue, X.: Lattice: Democratize high-fidelity 3d generation at scale. arXiv preprint arXiv:2512.03052 (2025)
27. Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., Bi, S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. arXiv preprint arXiv:2311.06214 (2023)
28. Li, P., Liu, Y., Long, X., Zhang, F., Lin, C., Li, M., Qi, X., Zhang, S., Luo, W., Tan, P., et al.: Era3d: High-resolution multiview diffusion using efficient row-wise attention. arXiv preprint arXiv:2405.11616 (2024)
29. Li, P., Ma, S., Chen, J., Liu, Y., Zhang, C., Xue, W., Luo, W., Sheffer, A., Wang, W., Guo, Y.: Cmd: Controllable multiview diffusion for 3d editing and progressive generation. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–10 (2025)
30. Li, P., Zheng, W., Liu, Y., Yu, T., Li, Y., Qi, X., Chi, X., Xia, S., Cao, Y.P., Xue, W., et al.: Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In: Proceedings of the computer vision and pattern recognition conference. pp. 16008–16018 (2025)

31. Li, W., Liu, J., Chen, R., Liang, Y., Chen, X., Tan, P., Long, X.: Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. arXiv preprint arXiv:2405.14979 (2024)
32. Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W., et al.: Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. arXiv preprint arXiv:2502.06608 (2025)
33. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR. pp. 300–309 (2023)
34. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: ICCV. pp. 5741–5751 (2021)
35. Liu, G., Yang, B., Zhi, Y., Zhong, Z., Ke, L., Deng, D., Gao, H., Huang, Y., Zhang, K., Fu, H., et al.: Beyond vlm-based rewards: Diffusion-native latent reward modeling. arXiv preprint arXiv:2602.11146 (2026)
36. Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J., Su, H.: One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In: CVPR. pp. 10072–10083 (2024)
37. Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. NeurIPS **36** (2023)
38. Liu, Y., Xie, M., Liu, H., Wong, T.T.: Text-guided texturing by synchronized multi-view diffusion. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
39. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
40. Luo, D., Yang, S., Yang, M., Lu, J., Tang, Y., Han, X., Chen, Z., Wang, B., Guo, C.: Matpedia: A universal generative foundation for high-fidelity material synthesis. arXiv preprint arXiv:2511.16957 (2025)
41. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: CVPR. pp. 2837–2845 (2021)
42. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV. pp. 405–421 (2020)
43. Müller, N., Siddiqui, Y., Porzi, L., Bulo, S.R., Kontschieder, P., Nießner, M.: Diffrf: Rendering-guided 3d radiance field diffusion. In: CVPR. pp. 4328–4338 (2023)
44. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
45. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal pp. 1–31 (2024)
46. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
47. Qiu, L., Chen, G., Gu, X., Zuo, Q., Xu, M., Wu, Y., Yuan, W., Dong, Z., Bo, L., Han, X.: Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In: CVPR. pp. 9914–9925 (2024)
48. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
49. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: ECCV. pp. 501–518. Springer (2016)

50. Shue, J.R., Chan, E.R., Po, R., Ankner, Z., Wu, J., Wetzstein, G.: 3d neural field generation using triplane diffusion. In: CVPR. pp. 20875–20886 (2023)
51. Sun, H., Gao, Y., Xie, J., Yang, J., Wang, B.: Svg-ir: Spatially-varying gaussian splatting for inverse rendering. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16143–16152 (2025). https://doi.org/10.1109/CVPR52734.2025.01505
52. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: ECCV. pp. 1–18. Springer (2025)
53. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In: ICLR (2024)
54. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In: ICCV. pp. 22819–22829 (2023)
55. Tang, Z., Gu, S., Wang, C., Zhang, T., Bao, J., Chen, D., Guo, B.: Volumediffusion: Flexible text-to-3d generation with efficient volumetric encoder. arXiv preprint arXiv:2312.11459 (2023)
56. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
57. Tencent Hunyuan: Hunyuan3d. https://3d.hunyuan.tencent.com/ (2024)
58. Tripo AI: Tripo: Fast 3d object generation from text and image (2024), https://www.tripo3d.ai/, accessed: 2024-05-20
59. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: Patchmatchnet: Learned multi-view patchmatch stereo. In: CVPR. pp. 14194–14203 (2021)
60. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. arXiv preprint arXiv:2503.11651 (2025)
61. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR. pp. 20697–20709 (2024)
62. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: CVPR. pp. 4563–4573 (2023)
63. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. NeurIPS **36** (2024)
64. Wang, Z., Wang, Y., Chen, Y., Xiang, C., Chen, S., Yu, D., Li, C., Su, H., Zhu, J.: Crm: Single image to 3d textured mesh with convolutional reconstruction model. In: ECCV. pp. 57–74. Springer (2025)
65. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: Nerf–: Neural radiance fields without known camera parameters (2021)
66. Wei, X., Zhang, K., Bi, S., Tan, H., Luan, F., Deschaintre, V., Sunkavalli, K., Su, H., Xu, Z.: Meshlrm: Large reconstruction model for high-quality mesh. arXiv preprint arXiv:2404.12385 (2024)
67. Wu, C.H., Chen, Y.C., Solarte, B., Yuan, L., Sun, M.: ifusion: Inverting diffusion for pose-free reconstruction from sparse views. arXiv preprint arXiv:2312.17250 (2023)
68. Wu, K., Liu, F., Cai, Z., Yan, R., Wang, H., Hu, Y., Duan, Y., Ma, K.: Unique3d: High-quality and efficient 3d mesh generation from a single image. arXiv preprint arXiv:2405.20343 (2024)

69. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Xu, J., Torr, P., Cao, X., Yao, Y.: Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. arXiv preprint arXiv:2405.14832 (2024)
70. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., Yang, J.: Native and compact structured latents for 3d generation (2025), https://arxiv.org/abs/2512.14692
71. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. arXiv preprint arXiv:2412.01506 (2024)
72. Xu, C., Li, A., Chen, L., Liu, Y., Shi, R., Su, H., Liu, M.: Sparp: Fast 3d object reconstruction and pose estimation from sparse views. In: ECCV. pp. 143–163. Springer (2024)
73. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024)
74. Xu, Q., Tao, W.: Multi-scale geometric consistency guided multi-view stereo. In: CVPR. pp. 5483–5492 (2019)
75. Xu, Y., Shi, Z., Yifan, W., Chen, H., Yang, C., Peng, S., Shen, Y., Wetzstein, G.: Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. arXiv preprint arXiv:2403.14621 (2024)
76. Xu, Y., Tan, H., Luan, F., Bi, S., Wang, P., Li, J., Shi, Z., Sunkavalli, K., Wetzstein, G., Xu, Z., et al.: Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. In: ICLR (2024)
77. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1179–1189 (June 2023)
78. Yang, J., Mao, W., Alvarez, J.M., Liu, M.: Cost volume pyramid based depth inference for multi-view stereo. In: CVPR. pp. 4877–4886 (2020)
79. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: ECCV. pp. 767–783 (2018)
80. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent mvsnet for high-resolution multi-view stereo depth inference. In: CVPR. pp. 5525–5534 (2019)
81. Ye, C., Wu, Y., Lu, Z., Chang, J., Guo, X., Zhou, J., Zhao, H., Han, X.: Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. arXiv preprint arXiv:2503.22236 **3** (2025)
82. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. arXiv preprint arXiv:2506.04216 (2025)
83. Ye, Z., Liu, Q., Wei, C., Zhang, Y., Wang, X., Wan, P., Gai, K., Luo, W.: Visual-aware cot: Achieving high-fidelity visual consistency in unified models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9116–9126 (2026)
84. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. TOG **42**(4), 1–16 (2023)
85. Zhang, B., Cheng, Y., Yang, J., Wang, C., Zhao, F., Tang, Y., Chen, D., Guo, B.: Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. arXiv e-prints pp. arXiv–2403 (2024)

86. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. TOG **43**(4), 1–20 (2024)
87. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. arXiv preprint arXiv:2501.12202 (2025)
88. Zhao, Z., Liu, W., Chen, X., Zeng, X., Wang, R., Cheng, P., Fu, B., Chen, T., Yu, G., Gao, S.: Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. NeurIPS **36** (2024)
89. Zhou, J., Wang, J., Ma, B., Liu, Y.S., Huang, T., Wang, X.: Uni3d: Exploring unified 3d representation at scale. In: International Conference on Learning Representations (ICLR) (2024)
90. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: ICCV. pp. 5826–5835 (2021)
91. Zuo, Q., Gu, X., Qiu, L., Dong, Y., Zhao, Z., Yuan, W., Peng, R., Zhu, S., Dong, Z., Bo, L., et al.: Videomv: Consistent multi-view generation based on large video generative model. arXiv preprint arXiv:2403.12010 (2024)