ROAD: Reciprocal-Objective Alignment of Discriminative Semantics for 3D Shape Generation

Xiao Luo[1] Mingyang Du[1] Xin Zhou[1] Tianrui Feng[1]
Xiwu Chen[2] Xiaofan Li[3] Jiangning Zhang[3] Dingkang Liang[1*]
[1]Huazhong University of Science and Technology, China
[2]Megvii, China [3]Zhejiang University, China
{dkliang, tianruifeng, xzhou03}@hust.edu.cn
*Corresponding author

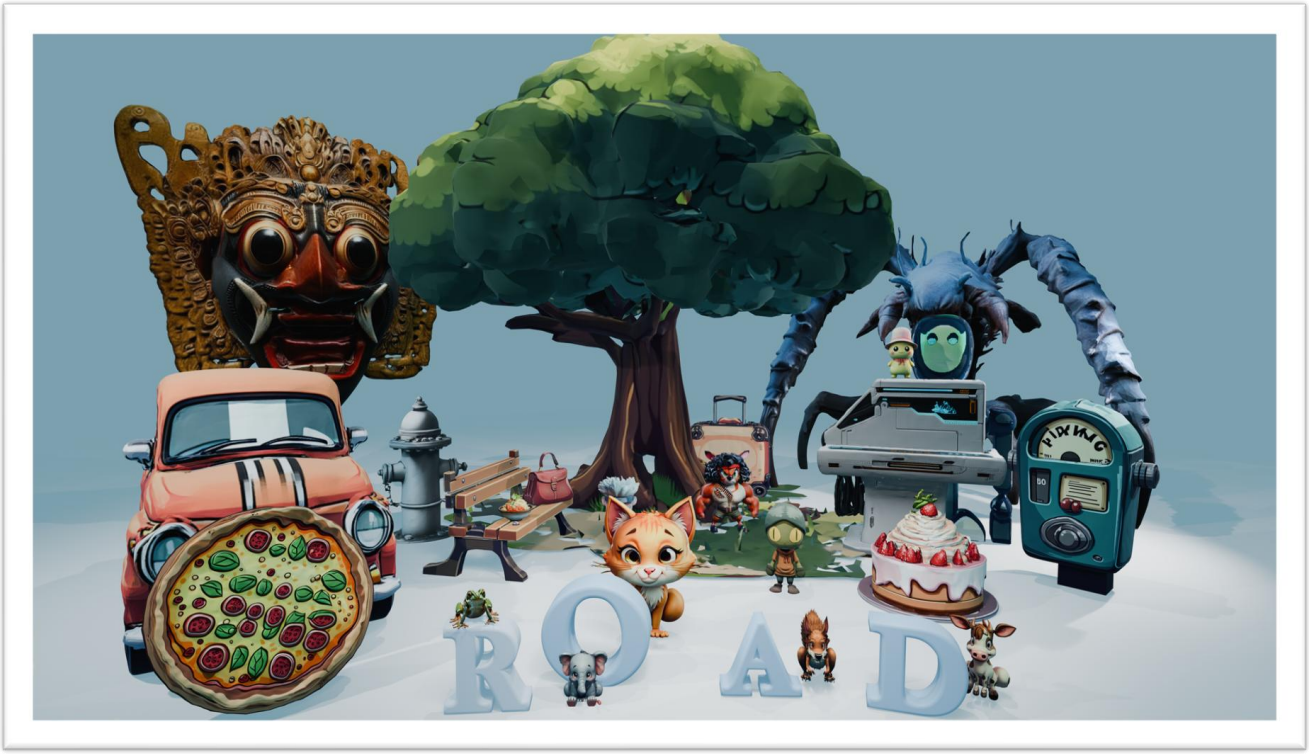


Figure 1. High-fidelity 3D assets generated by ROAD. These results showcase the capability of our method to synthesize diverse 3D models across various categories and styles with superior geometric and textural quality.

Abstract

High-fidelity 3D generation predominantly relies on scaling model capacity and data, which incurs prohibitive computational costs. This paradigm typically requires learning geometry from scratch and overlooks the rich semantic and structural priors already encapsulated in discriminative 3D foundation models. We contend that leveraging the profound understanding of the 3D world possessed by these discriminative models can significantly reduce generative cost. To this end, we propose ***ROAD****, a framework that reduces the training cost of 3D generation by transferring these rich discriminative priors into diffusion transformers. To address the inherent semantic-structural heterogeneity between generative and discriminative latents, we introduce a reciprocal-objective alignment strategy. This method synergizes Holistic Semantic Condensing to enforce global semantic coherence and Structural Optimal Alignment which*

*is formulated as a bipartite matching problem to rigorously align microscopic geometric details between disparate latent spaces. The 3D foundation model is only used for training-time supervision of alignment and is not used at inference, incurring no additional inference cost. Compared with industrial baseline Step1X-3D, the proposed **ROAD** achieves highly competitive generation performance with only 1.5% of the training data and significantly reduces training costs, effectively reducing the computational overhead of high-fidelity 3D generation. The code will be made available.*

1. Introduction

3D Shape Diffusion Models (3SDMs) have emerged as the dominant paradigm for generating high-fidelity 3D assets. Recent frontiers, such as Hunyuan3D 2.1 [9], Step1X-3D [15], and TRELLIS [51], have pushed the boundaries of geometric realism. However, their success relies heavily on a scaling strategy, e.g., Step1X-3D consumes 96 $\times$ A100 GPUs over 4 days on 2,000k data. We argue that this paradigm masks a fundamental cognitive inefficiency: these models attempt to learn to generate 3D geometry from scratch, which requires massive resources to establish the mapping between conditions and geometric structures, ignoring the fact that discriminative 3D Foundation Models (3DFMs) [19, 22, 52, 53, 62] have encapsulated a profound understanding of the 3D world. Why compel a generative model to relearn the world from raw geometric primitives, when it could inherit the rich, pre-trained structural priors from a discriminative foundation model?

Recently, using pretrained discriminative foundation models to guide generative training has become an active direction in both image [13] and video [60] generation. By properly integrating the priors of foundation models from their respective modalities, these works have shown remarkable improvements in generative quality and capability. Therefore, we believe that leveraging the large-scale pre-trained knowledge of 3DFMs to bootstrap 3SDMs is a promising direction. Given that these foundation models have acquired universal 3D discriminative representations from billion-scale training, they fundamentally capture the essential features of diverse 3D geometries at the latent level. Injecting these strong priors acts as a cognitive shortcut, potentially allowing shape generative models to achieve high-fidelity synthesis with a fraction of the data.

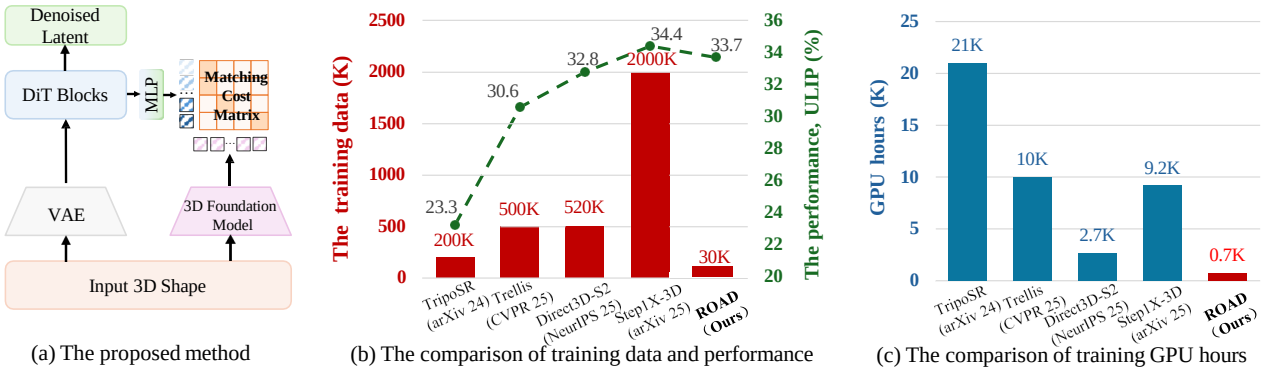
However, bridging the gap between discriminative priors and generative latents is non-trivial due to the inherent semantic-spatial misalignment. In the domains of image and video generation, a naive or intuitive approach is often adopted to enforce spatial element-wise alignment (i.e., assuming strict positional correspondence between tokens). This strategy proves effective primarily due to the high structural similarity of their encoding frameworks. Specifically, both discriminative and generative models typically process the inputs using identical mechanisms such as shared grid-based patch partitioning, inherently preserving consistent spatial processing relationships and maintaining a strictly aligned order within the token. Yet, our investigation reveals that this strategy is fundamentally flawed in 3D generation due to semantic-structural heterogeneity.

We attribute this to two factors: first, the permutation-invariant nature of 3D point clouds essentially leads to distinct patch groupings, making deterministic token alignment nearly intractable. Second, more importantly, 3D foundation models and generative models rely on disparate encoding architectures. Generative models typically aggregate global surface context into sparse latent anchors, with each token attempting to represent the overall surface information, resulting in tokens that are holistic and position-agnostic. In contrast, 3DFMs tend to encode grouped localized geometric features that maintain strong regional distribution. Consequently, even when processing the identical spatial patch, 3SDMs and 3DFMs often represent entirely disparate information for one encompassing global context and the other capturing local details. This structural-semantic discrepancy renders standard spatial element-wise alignment ineffective, as there is no natural one-to-one correspondence between the two latent sequences.

To unlock the potential of representation alignment in 3D shape generation, we strive to seek a simple yet versatile approach that bridges this inherent mismatch between generative and foundation models. While directly integrating their features to align decoupled high-level semantics might seem like a straightforward solution, it inevitably suffers from a lack of fine-grained constraints, for generative models heavily rely on local details. Given that each token in both models inherently encapsulates geometric information, naturally, we argue that establishing explicit correspondences between these two unordered sets is of paramount importance. The successful application of Hungarian Matching to address similar set-to-set assignment problems in 2D object detection [1] serves as a profound inspiration for our approach.

In this paper, we propose **ROAD** (**R**eciprocal **O**bjective **A**lignment of **D**iscriminative Semantics for 3D Shape Generation), a plug-and-play framework designed to infuse geometric representation of 3DFMs into 3SDMs. Our framework harmonizes the disparate latents and establishes a multi-granularity synergistic alignment, enabling the generative model to simultaneously assimilate macroscopic semantic abstractions and microscopic geometric primitives while preserving its native flexibility.

Specifically, this mechanism efficiently enforces representation coherence through two synergistic pathways: Holistic Semantic Condensing (HSC) and Structural Opti-



(a) The proposed method (b) The comparison of training data and performance (c) The comparison of training GPU hours

Figure 2. (a) Overview of our proposed ROAD framework. Instead of relying on resource-intensive, brute-force scaling, we introduce a highly streamlined design tailored for semantic-structural alignment. (b) Computation-performance Pareto frontier. We benchmark our method against prior state-of-the-art works, showcasing a standard-setting trade-off between training data requirements and downstream efficacy. (c) Unprecedented training efficiency. Our method requires the absolute minimum GPU hours among all existing approaches under equivalent or estimated hardware configurations.

mal Alignment (SOA). The HSC employs global anchoring to aggregate disordered tokens into a semantic centroid, compelling the generated latent distribution to align with the macroscopic semantic manifold defined by the foundation model. In contrast, SOA formulates token correspondence as a bipartite matching problem, leveraging the Hungarian algorithm to dynamically match tokens based on semantic similarity for granular alignment. These two reciprocal objectives ensure the effective transfer of features, without disrupting the native generative flow.

We evaluate our approach using the representative open-source framework Step1X-3D [15]. We find that ROAD substantially reduces the computational cost, as shown in Fig. 2(b)-(c). By acting as a concise semantic bridge, ROAD achieves performance with significantly fewer parameters and training resources, outperforming several state-of-the-art models. These results demonstrate that semantic alignment is a potent substitute for massive data scaling, offering a path to a more accessible 3D generative community.

Our major contributions are as follows: **1)** We reveal the inherent representation misalignment between 3D foundation models and generative models, identifying this structural discrepancy as the primary obstacle to effective prior transfer. **2)** We propose ROAD, a framework that bridges these disparate latent spaces via a stratified alignment mechanism across both semantic abstraction and structural detail, ensuring rigorous prior injection. **3)** We demonstrate that effective prior transfer significantly reduces data and training costs while maintaining competitive fidelity, thereby realizing the efficiency of high-quality 3D generation.

2. Related Work

2.1. 3D Shape Generation

Recently, 3D shape generation, a core task in computer vision, has emerged as a prominent research direction, as it demonstrates the capacity to liberate 3D modeling from labor-intensive manual techniques through automated generation. However, various technical limitations persist. To lift lower-dimensional conditions to global 3D features, early optimization-based frameworks [26, 32, 43, 45, 54, 56] construct latent 3D representations from multi-view images, which are constrained by slow convergence and high rendering costs. Later works transitioned to efficient feed-forward networks trained on diverse representations, including point clouds [11, 16, 23, 39, 63], meshes [2, 34, 40], and hybrid features [8, 33, 38]. Fine-grained point-cloud generation and enhancement have also been explored through hierarchical graph modeling and frequency-aware generative learning [16, 21], further demonstrating the importance of preserving local geometric details in 3D synthesis. However, these models do not alleviate the reliance on extensive multi-view images, still imposing a heavy training burden.

Modern approaches typically employ Variational Autoencoders (VAEs) to compress geometry into compact latent spaces. This strategy successfully aggregates high-dimensional features, allowing for the derivation of full 3D shape structures from single-view inputs. While state-of-the-art models like Tripo series [17, 38], Hunyuan3D [9, 12], Step1X-3D [15], and TRELLIS [51] achieve remarkable fidelity, they nevertheless primarily rely on expanding model capacity and data, which incurs substantial computational demands and resource overhead, making them increasingly inaccessible to the broader academic community.

Unlike these approaches that rely on brute-force scaling,

we investigate empowering generative frameworks by transferring rich priors from discriminative 3D Foundation Models (3DFMs), achieving high-quality generation with significantly reduced computational overhead.

2.2. 3D Foundation Models

3D foundation models extract geometric and semantic information from unordered point clouds. Early approaches use shared MLPs for independent feature extraction [27, 30, 31], whereas later advancements leverage local structural contexts, inter-region relations, high-order geometric dependencies, and global representations [4, 20, 36, 41, 59]. These pioneering models establish a robust paradigm for processing unstructured 3D data, effectively transforming raw point clouds into high-dimensional latent representations enriched with deep semantic cues.

Significant strides have been made in 3D representation learning, driven by diverse architectures including the Point Transformer series [49, 50, 61] and PointMamba [18], as well as self-supervised frameworks like PointMAE [25], Recon [28], and masked point-cloud representation learning [42]. Moreover, cross-modal models such as PointCLIP [58] and ShapeLLM [29] have successfully scaled discriminative capabilities by aligning 3D data with rich semantic spaces. These models leverage massive datasets to encapsulate robust structural and semantic abstractions. Instead of limiting 3DFMs to discrimination, we propose transferring their rich semantic and structural priors (e.g., Uni3D [62]) to generative frameworks.

Despite their potential, the direct integration of strong discriminative features from 3D models poses significant hurdles [7]. ROAD bridges this representational mismatch, enabling the seamless injection of latent features into generative models. By aligning generative latents with these robust features, we bootstrap the generation process, significantly enhancing fidelity.

2.3. Representation Alignment

Recently, visual representation alignment (e.g., REPA [57], U-REPA [37], REPA-E [13]) has successfully enhanced the semantic information of image generative models by injecting representation information of foundation models, token by token, significantly boosting the training efficiency and convergence speed of generative models. Subsequent work SRA [10] replaces representation alignment with self-alignment. IREPA [35], and REG [46], enrich semantic guidance to refine REPA, while VideoREPA [60] and GeometryForcing [47] extend these methods to video generation. However, these approaches require a strict spatial, element-wise, one-to-one correspondence across the encoded tokens. In 3D shape generation, the semantic-structural heterogeneity between 3SDMs and 3DFMs disrupts the correlation, which presents an alignment challenge.

Consequently, in this paper, we propose a novel framework, ROAD, to explore representation alignment within generative models for 3D generation, thereby lowering the entry barriers to high-quality 3D content creation.

3. Preliminary

3.1. 3D Foundation Model Paradigm

To handle unordered point clouds, an input point cloud $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{N_p}$ with optional point attributes $\mathbf{A} = \{\mathbf{a}_i\}_{i=1}^{N_p}$ is first partitioned into G localized patches. For the g-th patch centered at $\mathbf{c}_g \in \mathbb{R}^3$, its local geometric encoding is formulated as:

$$f_g = \mathrm{Pool}_{\mathbf{q}_i \in \mathcal{N}(\mathbf{c}_g)} \left(\mathrm{MLP} \left([\mathbf{q}_i - \mathbf{c}_g; \mathbf{a}_i] \right) \right), \tag{1}$$

where $\mathcal{N}(\mathbf{c}_g)$ denotes the neighboring points around $\mathbf{c}_g$, and $\mathbf{a}_i$ represents the attribute of point i, such as normals or colors. The operator $[\cdot;\cdot]$ denotes feature concatenation.

The patch features are then projected and organized into a transformer input sequence:

$$\mathbf{F}^0 = \left[\mathbf{f}_{\mathrm{cls}} + \mathbf{p}_{\mathrm{cls}}; \{\mathrm{Proj}(f_g) + \mathrm{PosEnc}(\mathbf{c}_g)\}_{g=1}^{G} \right], \tag{2}$$

where $\mathbf{f}_{\mathrm{cls}}$ and $\mathbf{p}_{\mathrm{cls}}$ denote the learnable classification token and its positional embedding, respectively. $\mathrm{Proj}(\cdot)$ is a linear projection layer, and $\mathrm{PosEnc}(\cdot)$ denotes the positional encoding derived from patch centers. After the transformer encoder, we denote the extracted patch-level discriminative features as

$$\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{M} \in \mathbb{R}^{M \times D'}, \tag{3}$$

which serve as the 3D foundation-model priors used in our framework. Since point-cloud patches are produced by sampling and grouping operations, $\mathbf{Y}$ is treated as an unordered set of local semantic-structural tokens rather than a sequence with canonical ordering.

3.2. Hungarian Matching Paradigm

Given two unordered sets $\mathcal{U} = \{\mathbf{u}_i\}_{i=1}^{K}$ and $\mathcal{V} = \{\mathbf{v}_j\}_{j=1}^{K}$, Hungarian Matching provides a standard solution for establishing one-to-one correspondences without relying on predefined element order. For each pair $(\mathbf{u}_i, \mathbf{v}_j)$, a matching cost $\mathcal{D}_{ij}$ is computed according to the task requirement. The general formulation can be written as:

$$\mathcal{D}_{ij} = \lambda_{\mathrm{cls}} \mathcal{L}_{\mathrm{cls}}(\mathbf{u}_i, \mathbf{v}_j) + \lambda_{\mathrm{reg}} \mathcal{L}_{\mathrm{reg}}(\mathbf{u}_i, \mathbf{v}_j), \tag{4}$$

where $\mathcal{L}_{\mathrm{cls}}$ and $\mathcal{L}_{\mathrm{reg}}$ denote the classification and regression costs, and λ_{cls} and λ_{reg} are balancing coefficients.

The optimal assignment is obtained by searching for a permutation σ^* that minimizes the total matching cost:

$$\sigma^* = \underset{\sigma \in \mathfrak{S}_K}{\mathrm{argmin}} \sum_{i=1}^{K} \mathcal{D}_{i,\sigma(i)}, \tag{5}$$

where $\mathfrak{S}_K$ denotes the set of all permutations over K elements and $\sigma(i)$ denotes the mapped value for i-th input. After obtaining σ^*, the matched loss is computed as:

$$\mathcal{L}_{\text{match}} = \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}\left(\mathbf{u}_i, \mathbf{v}_{\sigma^*(i)}\right). \tag{6}$$

In this way, Hungarian Matching removes the ambiguity caused by unordered elements and enables permutation-invariant set-to-set supervision. In our method, we instantiate this paradigm with feature-similarity costs to align generative tokens with 3D foundation-model tokens.

4. Our Method

While representation alignment significantly bolsters the efficacy of generative models, a straightforward adaptation of successful 2D methodologies fails to adequately address the specialized requirements of 3D synthesis. In this section, we present **ROAD**, a framework that scales down the cost of high-fidelity generation by injecting the rich priors of discriminative 3D Foundation Models (3DFMs) into 3D Shape Diffusion Models (3SDMs). As illustrated in Fig. 3, our framework operates by integrating frozen 3DFMs (e.g., Uni3D) as a semantic supervisor alongside the standard generative framework. Here, the frozen 3DFM is employed only during training to compute the proposed alignment constraints. To inject discriminative priors without compromising the generative model's flexibility, we introduce a reciprocal-objective alignment stream comprising two synergistic components: Holistic Semantic Condensing (HSC) for global semantic coherence, and Structural Optimal Alignment (SOA) for fine-grained geometric fidelity. By employing this dual-alignment strategy, the potent discriminative priors from the foundation model are effectively integrated into the generative process. Consequently, this approach significantly reduces training overhead while enabling the synthesis of high-quality 3D assets.

4.1. Overall Framework

To evaluate the feasibility of our approach, we adopt the state-of-the-art Step1X-3D [15] framework as our baseline. This framework generally comprises a SDF Variational Autoencoder (VAE) for geometric compression and a Diffusion Transformer (DiT) for latent generation.

SDF Variational Autoencoder. To compress continuous geometry into a compact representation, the framework utilizes an SDF-VAE [3]. Given a point set $\mathcal{P}$, the encoder E_ϕ employs a cross-attention to aggregate global context onto N sparse spatial anchors sampled (key points) via Farthest Point Sampling (FPS), yielding latent tokens $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N} \in \mathbb{R}^{N \times D}$, where D denotes the feature dimension. A decoder D_θ then queries these tokens to predict the Signed Distance Field (SDF) values for surface reconstruction. Such an SDF-VAE compresses complex geometric spaces into dense SDF fields, where each token processes global features rather than localized aggregation.

3D Diffusion Transformer. The generative backbone is a Multi-modal Diffusion Transformer (MM-DiT) [6] designed to model the latent distribution, which is trained under a flow-matching paradigm. Operating directly on $\mathbf{X}$, the model learns a velocity field over continuous latent trajectories conditioned on semantic context $\mathbf{c}$. Specifically, given a clean latent $\mathbf{x}_0$ encoded by the VAE and a Gaussian noise latent $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we construct an interpolated latent

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad t \in (0,1), \tag{7}$$

and train the network to predict the corresponding flow velocity. The architecture incorporates double-stream and single-stream blocks to facilitate interaction between geometric tokens and condition features, optimized via the diffusion objective:

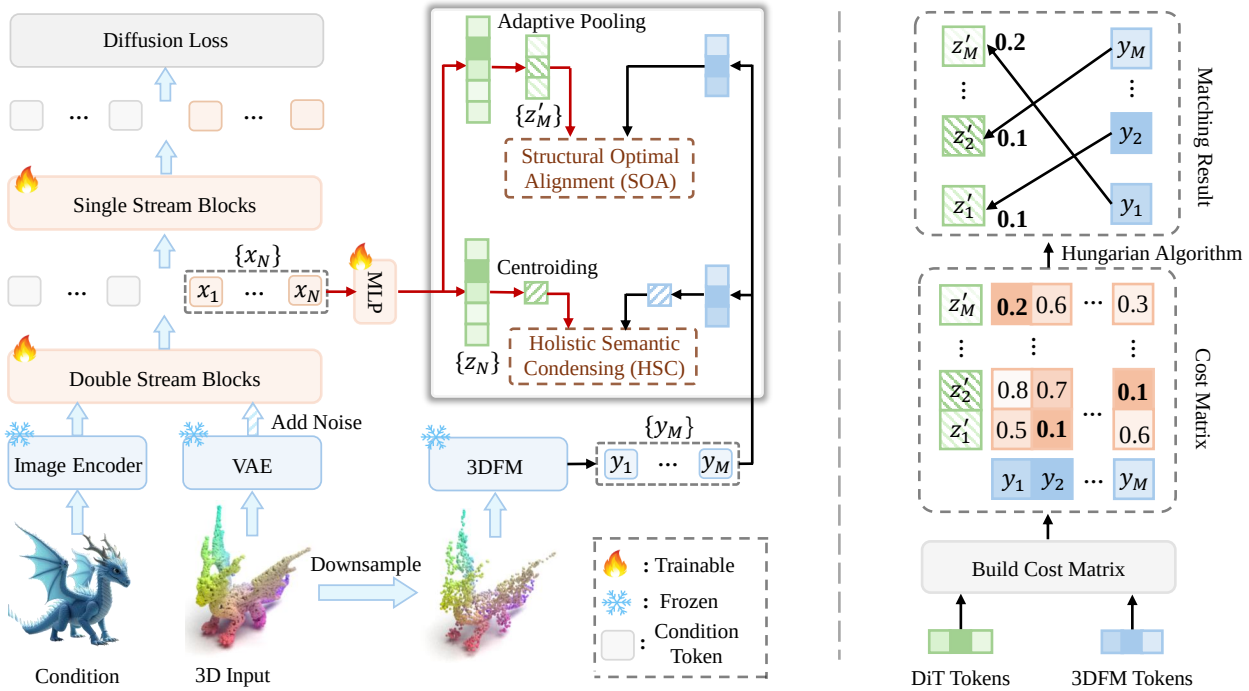
$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1, t, \mathbf{c}} \left[\left\| (\mathbf{x}_1 - \mathbf{x}_0) - G_\theta(\mathbf{x}_t, t, \mathbf{c}) \right\|^2 \right]. \tag{8}$$

In the overall training pipeline, the VAE first encodes the 3D shape into latent tokens, which are then linearly interpolated with Gaussian noise and fed into the DiT alongside conditional inputs to learn the latent flow. Our ROAD framework intervenes in this flow by injecting discriminative priors from a frozen 3D foundation model into DiT blocks via the proposed reciprocal-objective constraints.

4.2. Investigation: Semantic-Structural Heterogeneity

Before elaborating on the dual-alignment methodology, we first identify the fundamental barrier hindering the direct transfer of priors: the semantic-structural heterogeneity between generative and discriminative representations. To fully characterize this misalignment, it is essential to analyze the inherent disorder of point clouds alongside the structural discrepancies between these two representational paradigms.

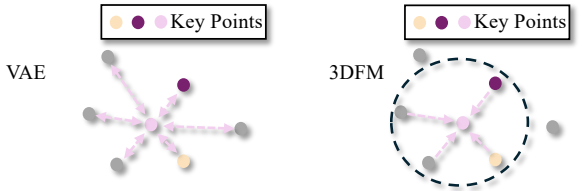
For densely sampled point clouds, existing methods predominantly rely on random sampling or Farthest Point Sampling (FPS) to select cluster centers, which are then organized into token sequences. However, this approach inherently introduces stochasticity into the sequence composition. Consequently, in tasks requiring precise token-to-token alignment, 3D representations lack a standardized serialization protocol, unlike the well-defined grid-based tokenization prevalent in image preprocessing. A straightforward approach would be to forcibly fix the point sequences of the generative model with those of the foundation model. However, this strategy encounters a subsequent challenge, as the two paradigms employ fundamentally different encoding strategies for point representation.



(a) The overview of ROAD (b) The detail of SOA

Figure 3. Overview of the proposed ROAD. (a) During training, the frozen 3D Foundation Model (3DFM) provides semantic priors to guide the diffusion process through complementary holistic consistency and similarity-based semantic alignment. (b) Details of Structural Optimal Alignment (SOA). We build a cost matrix between DiT tokens and 3DFM tokens, and then apply the Hungarian algorithm to obtain token-level matching results for semantic alignment.

Initially, we observe that the feature sequence encoded by the SDF-VAE exhibits permutation invariance. Specifically, the decoder can reconstruct the correct geometry even if the latent tokens are shuffled. This implies that the VAE hidden states do not strictly encode sequential correspondence. We attribute this to the tokenization mechanism shown in Fig. 4(a), where the VAE employs cross-attention to aggregate context from the entire point set $\mathcal{P}$ onto sparse anchors, whereas standard 3DFMs typically utilize explicit grouping operations (e.g., KNN) to encode geometry from local features. Consequently, DiT tokens behave as order-agnostic latent descriptors of the surface rather than strictly localized patch descriptors, yet 3DFMs' tokens maintain strong spatial correlation. This discrepancy is indicated by the feature visualization in Fig. 4(b), where points with similar color encode homogeneous information. The visualization confirms that the generative sequence lacks a fixed spatial correspondence with the discriminative sequence, rendering rigid element-wise alignment fruitless and necessitating our proposed dual-alignment strategy.



(a) The different tokenization methods of VAE and 3DFM.

VAE: Global Context 3DFM: Local Feature

(b) The different semantic representations of VAE and 3DFM.

Figure 4. Analysis of semantic misalignment. (a) A showcase of different tokenization methods. (b) Feature projections confirm the representation gap.

4.3. Reciprocal-Objective Prior Alignment Mechanism

To resolve the semantic-structural heterogeneity identified above, we propose a unified mechanism that harmonizes representations at both the global and local levels. This process involves feature projection, holistic condensing, and structural bipartite matching.

Feature Projection and Normalization. Informed by

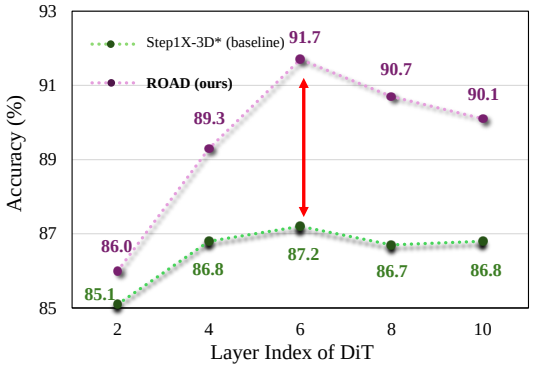


Figure 5. Classification result on ModelNet40. Our model significantly enhances the semantic information of the representations.

recent research in representation alignment, we begin by performing feature projection. Let $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{M} \in \mathbb{R}^{M \times D'}$ denote the features extracted from the frozen 3DFM, where M represents the number of latent tokens and D' denotes the feature dimensionality. We extract the intermediate features $\mathbf{X}$ from 3SDMs (see Sec. 4.4) and employ a lightweight MLP projector Φ_{MLP} to map $\mathbf{X}$ into the 3DFM's latent space, yielding $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{N} = \Phi_{\text{MLP}}(\mathbf{X}) \in \mathbb{R}^{N \times D'}$. To prevent numerical discrepancies from biasing the alignment, we apply L_2 normalization to yield $\hat{\mathbf{z}}_i$ and $\hat{\mathbf{y}}_j$. These preparations significantly enhance the stability of the alignment and encourage the model to focus on more informative features in subsequent alignment.

Holistic Semantic Condensing (HSC). Given the semantic-structural heterogeneity between the two models, direct element-wise alignment is unstable. Therefore, we first attempt to enforce alignment at the macroscopic level. We utilize Holistic Semantic Condensing to aggregate disordered tokens into unified semantic prototypes via global average pooling:

$$\mathbf{c}_z = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{z}}_i, \quad \mathbf{c}_y = \frac{1}{M}\sum_{j=1}^{M} \hat{\mathbf{y}}_j. \tag{9}$$

This holistic anchoring method resolves the unordered nature of 3D representations, which treats tokens as geometry integrals, overlooking the discrepancies between them. The HSC loss is then defined as the cosine distance between these centers, compelling the generative model to orient towards the foundation model's semantic centroids:

$$\mathcal{L}_{\text{HSC}} = 1 - \cos(\mathbf{c}_z, \mathbf{c}_y). \tag{10}$$

Structural Optimal Alignment (SOA). While HSC ensures global semantic coherence, the condensing process inevitably discards spatial details. Prior point-cloud generation and enhancement studies have shown that fine-grained geometric cues are crucial for preserving structural fidelity [16, 21]. To improve fine-grained structural information without rigid element-wise alignment, we formulate the local mapping as a bipartite matching problem. Specifically, we apply small-scale adaptive average pooling to $\mathbf{Z}$, obtaining $\mathbf{Z}' = \{\mathbf{z}'_i\}_{i=1}^{M} \in \mathbb{R}^{M \times D'}$. This step establishes the equal-cardinality condition required for one-to-one Hungarian matching, while also serving as a lightweight denoising operation to suppress training-induced perturbations in the latent tokens. Then, we compute a pairwise cost matrix $\mathbf{C} \in \mathbb{R}^{M \times M}$ based on the cosine distance between all token pairs:

$$\mathbf{C}_{i,j} = 1 - \cos(\mathbf{z}'_i, \mathbf{y}_j). \tag{11}$$

We seek an optimal permutation π^* from the set of all possible permutations $\{\pi\}$ that minimizes the total matching cost using the Hungarian algorithm:

$$\pi^* = \underset{\pi}{\text{argmin}} \sum_{i=1}^{M} \mathbf{C}_{i,\pi(i)}. \tag{12}$$

where $\pi(i)$ indicates the mapped index of the i-th element under the permutation π.

Consequently, following the mapping provided by π^*, we align the i-th token from 3SDM with the $\pi^*(i)$-th token of 3DFM using the same cosine distance as in $\mathbf{C}$. The SOA loss is then formulated as the average cosine distance between matching tokens:

$$\mathcal{L}_{\text{SOA}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{C}_{i,\pi^*(i)}. \tag{13}$$

With the implementation of SOA, the model is empowered to capture microscopic information missed in HSC while overcoming the limitations of semantic discrepancies in rigid element-wise alignment, thereby enforcing fine-grained local constraints and providing it with more concrete semantic details.

Training Objective. The final training objective combines the standard diffusion loss with our dual constraints:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_1 \mathcal{L}_{\text{HSC}} + \lambda_2 \mathcal{L}_{\text{SOA}}, \tag{14}$$

where $\lambda_1 = \lambda_2 = 0.5$ are weighting coefficients used to balance the contribution of each loss term.

4.4. Analysis on Semantic Layer Selection

Recent research indicates that varying alignment depths exert distinct influences on model performance. To identify the optimal layer for feature alignment that yields richer semantic priors, we investigate the generative backbone's internal representation hierarchy. Specifically, we utilize the Step1X-3D*[1] as our probe subject, extracting intermediate features from the frozen DiT blocks and training

[1] * *indicates that the Step1X-3D DiT is trained from scratch in our configuration (see Sec. 5.1)*

Table 1. Efficiency benchmarking against state-of-the-art 3D generative models. We compare ROAD with leading methods trained on both proprietary and public datasets. ROAD achieves highly competitive performance while utilizing substantially fewer training data (only 30k) and significantly fewer parameters, effectively facilitating high-fidelity 3D generation.

Model	Venue	Params.	Public Data	Private Data	Uni3D-Score ↑	ULIP-Score ↑
Commercial Models						
Hunyuan3D-2.1 [9]	arXiv 25	3.3B	100k	✓	33.3	33.1
Step1X-3D [15]	arXiv 25	1.3B	800k	1200k	33.5	34.4
Academic Models						
TripoSR [38]	arXiv 24	0.4B	200k	✗	24.6	23.3
TRELLIS [51]	CVPR 25	2.0B	500k	✗	26.8	30.6
Hi3DGen [55]	ICCV 25	4.2B	170k	700k	31.0	32.8
Direct3D-S2 [48]	NeurIPS 25	2.0B	520k	✗	31.8	32.8
ROAD (**ours**)	-	0.65B	30k	✗	**32.3**	**33.7**

lightweight linear classifiers on the ModelNet40 dataset. The results, visualized in Fig. 5, reveal that the 6^{th} layer reaches peak performance. Following this, accuracy gradually declines, suggesting a transition from semantic representation to abstract geometric encoding. Consequently, we explicitly target this layer for our alignment, ensuring that our discriminative priors are injected into the most receptive layers.

5. Experiment

5.1. Implementation details

We adopt the state-of-the-art 3D diffusion model, Step1X-3D [15], as our generative baseline. Note that to improve efficiency, we streamline the model architecture by reducing the parameter count by ∼50%, termed Step1X-3D*. Specifically, while the original Step1X-3D comprises 12 double-stream blocks and 24 single-stream blocks, our pruned variant Step1X-3D* utilizes 6 and 12, respectively. We employ the pre-trained Uni3D-G [62] as our discriminative foundation model for alignment. During training, we only update the DiT parameters and the projection MLP. This DiT is trained entirely from scratch without loading any pre-trained weights. Based on semantic analysis (see Sec. 4.4), we extract the intermediate features from the 6^{th} layer (double-stream) to perform the proposed alignment. All experiments are conducted on 8 NVIDIA A100 GPUs, with a per-GPU batch size of 16 and a learning rate of 1×10^{-3}. The model is trained for 130k iterations, taking approximately 3.5 days to converge.

5.2. Datasets and Evaluation Metrics

We construct our training set by only randomly sampling 30,000 assets from the publicly available Objaverse [5] dataset. The data processing pipeline strictly adheres to the protocols established by Step1X-3D. In a departure from the original setup, we render only 12 views for each 3D asset, as opposed to the 20 views utilized in the baseline.

Table 2. The comparisons between our method and other visual representation alignment methods.

Alignment methods	Venue	Uni3D-Score ↑	ULIP-Score ↑
Baseline	-	31.3	32.5
REPA [57]	ICLR 25	30.8	31.3
REG [46]	NeurIPS 25	30.1	28.2
SRA [10]	ICLR 26	29.4	30.1
ROAD (**ours**)	-	**32.3**	**33.7**

For evaluation, we curate a robust test set comprising 160 unseen assets, strictly ensuring asset-level non-overlap with the training data. Following previous methods [9, 15, 48], we assess the semantic consistency using standard CLIP-Score metrics, reporting results derived from Uni3D-G [62] and ULIP [52].

5.3. Main results

Comparison with state-of-the-art methods. We perform a comprehensive evaluation against a wide range of frontier 3D generative models, classifying them into commercial solutions and academic frameworks. As detailed in Table 1 and Fig. 2, ROAD demonstrates exceptional performance-efficiency trade-offs. Specifically, when compared to TripoSR [38], ROAD delivers substantial improvements in geometric quality and semantic consistency, outperforming it by significant margins in both Uni3D-Score (+7.7) and ULIP-Score (+10.4). Even against large-scale latent models such as TRELLIS [51] (2.0B params, 500k data), ROAD (0.65B, 30k) yields superior generative quality. Notably, regarding data efficiency, ROAD surpasses the recent Direct3D-S2 [48] (ULIP-Score: 32.8) while utilizing only 5.7% of its training data (30k vs. 520k) and one-third of its parameters, validating the economy of representation alignment in 3D shape generation.

Furthermore, ROAD challenges models reliant on massive proprietary datasets, including commercial giants like Step1X-3D [15] and Hunyuan3D [9], as well as the recent academic model Hi3DGen [55]. While these methods re-

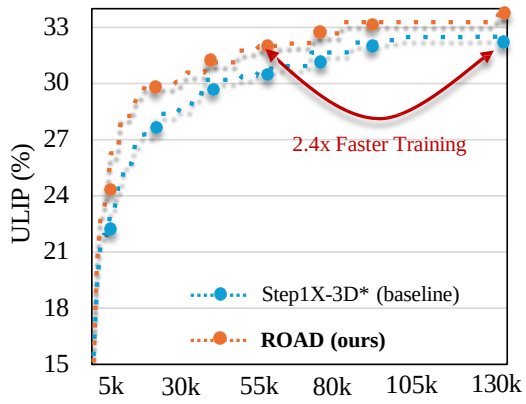


Figure 6. Convergence Analysis. Our method achieves the baseline's peak performance 2.4× faster (at 55k vs. 130k iterations).

quire million-scale private data (e.g., 700k–1200k) and prohibitive compute (e.g., 96 × GPUs), ROAD achieves parity with a mere public dataset and can be trained on a single node with just 8 A100 GPUs. By slashing resource requirements while maintaining exquisite quality, ROAD downscales high-fidelity 3D generation for the academic community.

Comparison with other representation alignment methods. To better elucidate the challenge of directly transferring 2D representation-alignment strategies to the 3D domain, we compare ROAD with recent visual representation alignment methods [10, 46, 57]. These methods are originally designed for image/video generation, where tokens usually follow regular grid structures or relatively stable positional correspondences. To adapt them to 3D generation, we further enforce exact coordinate correspondence between spatial point tokens to reduce the confounding randomness introduced by Farthest Point Sampling (FPS). As illustrated in Table 2, these transferred alignment strategies lead to noticeable performance degradation, whereas ROAD achieves consistent improvements. For rigid one-to-one alignment strategies such as REPA and REG, the enforced coordinate correspondence can disrupt the holistic surface representations of generative latents, resulting in inferior generation quality. For self-alignment strategies such as SRA, we conjecture that the relatively small generative model produces noisy intermediate representations during early training, making self-alignment less reliable before the model converges. In contrast, ROAD introduces a reciprocal-objective alignment paradigm that bridges the semantic-structural heterogeneity of 3D features through both holistic semantic guidance and similarity-based structural matching, effectively overcoming the limitations of rigid positional alignment.

Training convergence analysis. As illustrated in Fig. 6, the proposed ROAD demonstrates significantly faster convergence compared to the Step1X-3D* baseline. Quantitatively, our model reaches the baseline's peak performance (ULIP-Score $\approx$ 32.5%) at approximately 55k iterations, whereas the baseline requires over 130k iterations, effectively yielding a 2.4× acceleration. This speedup confirms that injecting discriminative priors effectively constrains the optimization search space, preventing the model from slowly learning semantic structures.

5.4. Qualitative Analysis

Generation fidelity compared with other models. As illustrated in Fig. 7 and Fig. 8, ROAD demonstrates improved generation quality compared to academic baselines (e.g., TRELLIS), delivering sharper geometric details and more faithful structures while mitigating over-smoothing artifacts. Most notably, ROAD achieves visual parity with industrial-scale models (e.g., Hunyuan3D), producing high-fidelity assets that are comparable in terms of structural integrity and realism, despite utilizing orders of magnitude less data. Furthermore, ROAD significantly rectifies the structural distortions and semantic inconsistencies inherent in the unaligned baseline, validating the feasibility of injecting discriminative priors and efficacy of our alignment strategy.

Visualizations of diverse 3D assets. To showcase the capability of our approach in generating high-fidelity 3D assets, we condition the generation process on a set of diverse input images collected from multiple sources. The final synthesized results are presented in Fig 9. Impressively, our model delivers exceptional generative performance across a wide range of object categories, thereby validating the robust generalization of ROAD.

5.5. Model Transferability

To verify the generalizability of our method, we also conduct experiments across diverse architectures, including **CraftsMan**[14] and **TRELLIS**[51]. For simplifying the experimental pipeline, unless specified otherwise, both models are trained and evaluated exclusively on the compact dataset declared in Sec. 5.2.

Transfer to Alternative SDF-Based Generators. Although CraftsMan shares a similar SDF latent representation with Step1X-3D, the native CraftsMan architecture eschews the sharp surface embedding employed by the latter. Furthermore, during the diffusion stage, CraftsMan relies on a conventional DiT backbone rather than the joint single-stream and dual-stream design found in Step1X-3D. Utilizing CraftsMan as a testbed, we investigate whether our proposed method maintains its generalizability under a framework constrained by less descriptive representations.

Specifically, we implement a straightforward alignment on the 6th DiT block of CraftsMan. Prior literature posits that early layers in generative models exhibit a stronger affinity for high-level semantic abstractions. Consequently,

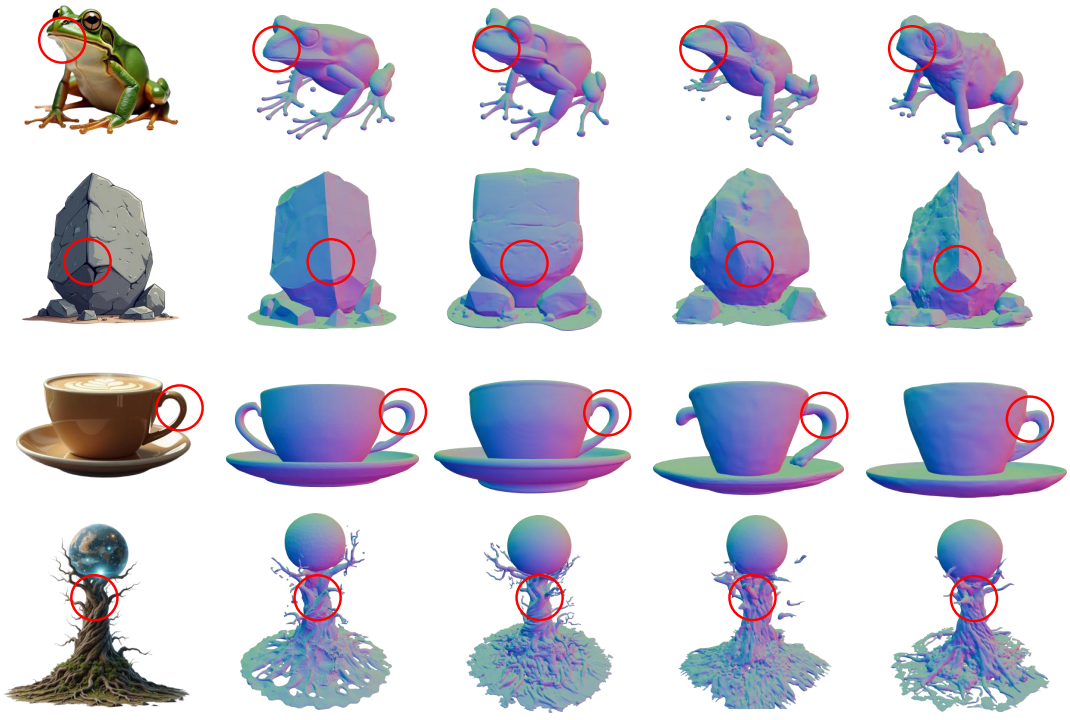


Figure 7. ROAD is comparable to TRELLIS (an academic benchmark) in geometric fidelity and achieves visual quality comparable to Hunyuan3D (an industrial paradigm), while effectively mitigating the structural artifacts observed in the baseline.

enforcing representational alignment at an early stage yields a more pronounced guidance effect. The quantitative results are detailed in Table 3. Our proposed method yields compelling results when integrated into the CraftsMan framework, successfully validating its applicability to fundamental generative baselines.

Generalization to Structured Sparse-Latent Architectures. TRELLIS adopts a novel structured latent representation to encode 3D geometric shapes. Specifically, this structured formulation voxelizes the underlying geometry, treating each occupied voxel block as an individual token. These tokens are subsequently organized via a specialized serialization protocol, ensuring that each token inherently encapsulates localized geometric information coupled with a distinct spatial part. However, this sparse voxel-based tokenization strategy still exhibits a pronounced discrepancy from the point-based tokenization paradigms utilized in foundation models. To validate the generalizability and robustness of our method with structured spatial latents, we apply our ROAD framework to TRELLIS. Concurrently, to streamline our training pipeline, we deliberately deploy our method within Stage 1, which exerts more profound impacts on geometric quality.

We strictly follow the method established in the original TRELLIS work to process our 30k dataset. As demonstrated by the quantitative evaluations compiled in Table 3, our proposed method is successfully validated on the TRELLIS framework. Specifically, the variant enhanced by our method consistently outperforms the baseline trained solely on the identical dataset without alignment. Furthermore, our model achieves competitive performance that is comparable to the original TRELLIS model trained under standard conditions. This consistent improvement demonstrates that ROAD can also benefit sparse structured latents and further underscores the cross-architecture stability of our framework.

Table 3. Results of CraftsMan and TRELLIS.

Variant	Uni3D-Score ↑	ULIP-Score ↑
Craftsman Model		
W/o ROAD	26.7	30.5
W/ ROAD	27.9	31.7
TRELLIS Model		
W/o ROAD	26.6	29.8
W/ ROAD	27.0	30.6

5.6. Ablation Studies

Unless otherwise stated, the following experiments are conducted on the Step1X-3D* baseline to validate the effectiveness of our method.

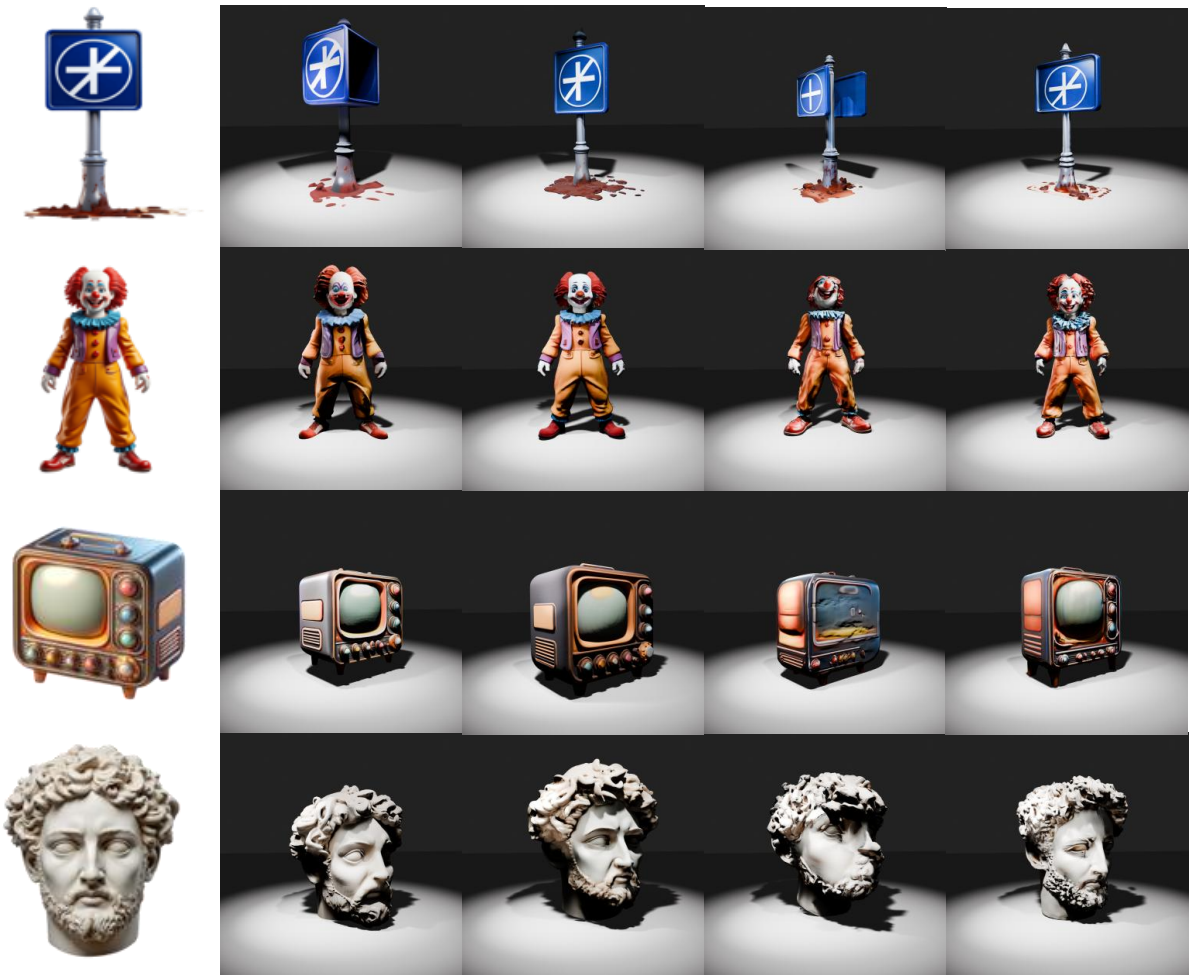


Figure 8. Our ROAD surpasses TRELLIS (academic models) in detail fidelity and achieves visual parity with Hunyuan3D (industrial models), while effectively rectifying the structural artifacts observed in the baseline.

The effect of reciprocal-objective alignment. We first verify the necessity of our dual-stream design. As presented in Table 4, deploying Holistic Semantic Condensing (HSC) or Structural Optimal Alignment (SOA) individually yields marginal improvements. This suggests that aligning solely at the macroscopic or microscopic level is insufficient to capture the foundation priors fully. However, their integration triggers a synergistic effect, boosting the ULIP-Score significantly to 33.7 (+1.2). We attribute this to their complementary nature: HSC acts as a global stabilizer, ensuring the generation aligns with the correct semantic manifold (what to generate), while SOA refines the local topology via bipartite matching (how to generate details). ROAD effectively harmonizes these two granularities.

Analysis on matching strategy: why semantics matter? A critical design choice in SOA is the metric used for bipartite matching. Table 5 reveals that employing spa-

Table 4. Impact of alignment components. We evaluate the contribution of Holistic Semantic Condensing (HSC) and Structural Optimal Alignment (SOA).

HSC	SOA	Uni3D-Score ↑	ULIP-Score ↑
-	-	31.3	32.5
✓	-	31.4	32.8
-	✓	31.7	33.2
✓	✓	**32.3**	**33.7**

Table 5. Effectiveness of different matching metrics for SOA. The experiments are equipped with HSC.

Basis	Matching Metric	Uni3D-Score ↑	ULIP-Score ↑
Spatial	Point Coordinate L_2	30.9	32.4
Semantic	Latent Feature L_2	31.6	33.0
	Latent Feature Cosine	**32.3**	**33.7**

Table 6. Impact of 3D Foundation Model (3DFM) Capacity.

Variant	Params	Uni3D-Score ↑	ULIP-Score ↑
W/o 3DFMs	-	31.3	32.5
W/ 3DFMs	0.3B	31.9	32.7
W/ 3DFMs	1B	**32.3**	**33.7**

Table 7. Exploring the Feasibility of Aligning Diverse Modalities.

Alignment modalities	Venue	Uni3D-Score ↑	ULIP-Score ↑
Baseline	-	31.3	32.5
DINOv2 [24]	TMLR 23	29.6	30.2
VGGT [44]	CVPR 25	30.4	30.0
Uni3D (**ours**)	-	**32.3**	**33.7**

tial matching based on keypoint coordinates leads to performance degradation. This empirical failure corroborates our hypothesis in Sec. 4.2 that tokens lack fixed spatial correspondence due to the permutation-invariant nature of the VAE. Consequently, forcing alignment based on physical coordinates disrupts the generative model's native latent topology. In contrast, Semantic Matching successfully bridges this gap by identifying correspondence based on feature content. Specifically, using Latent Feature Cosine similarity yields the optimal performance, verifying that ROAD aligns geometric details by respecting the semantic-structural heterogeneity between the two models.

Impact of foundation model capacity. We investigate the influence of the pre-trained 3D Foundation Model's (3DFM) scale on prior transfer effectiveness, as shown in Table 6. It can be observed that as the model size decreases, the performance gains introduced by alignment gradually diminish. This suggests that weaker discriminators may provide insufficient guidance, resulting in less stable intermediate representations in the generative model during training. Conversely, scaling up to larger foundation models leads to progressive improvements, confirming that a sufficiently strong discriminator is essential for capturing the rich semantic priors needed to guide the generative process.

Which modality of representational information is effective? To identify the optimal representational information for 3D Shape Diffusion Models (3SDMs), we explore foundation models from various modalities, including DINOv2, VGGT, and Uni3D (i.e., 2D, 2.5D, and 3D). As shown in Table 7, aligning with DINOv2 and VGGT leads to suboptimal performance. By contrast, our approach effectively aligns with native 3D Foundation Model (3DFM), obtaining significantly better results. This indicates that 3SDMs inherently rely on concrete geometry information. Consequently, ROAD elegantly injects the priors from the 3DFM, generating high-fidelity 3D assets. Specifically, 2D visual features mainly capture appearance-level semantics and lack a comprehensive understanding of complete 3D geometry, making them insufficient for guiding structured shape synthesis. While 2.5D representations provide stronger geometric cues, they are still derived from view-dependent observations and may not fully characterize the continuity and global topology of object surfaces. Consequently, forcibly injecting such mismatched priors can interfere with the generative capacity of 3SDMs. In contrast, native 3D foundation models directly encode spatial structures and surface-level geometric semantics, providing more compatible and informative priors for enhancing 3D generative modeling.

Table 8. Effect of Sampling Density.

Num of Points	Density	Uni3D-Score ↑	ULIP-Score ↑
-	-	31.3	32.5
4096	Small	31.9	33.4
10000	Middle	**32.3**	**33.7**
20000	Large	31.4	32.4

Table 9. Ablation study on different alignment layers.

Target Layer	Uni3D-Score ↑	ULIP-Score ↑
Shallow-to-Mid Range (Layers 1–6)		
Layer 2 (Double-Stream)	31.4	32.8
Layer 4 (Double-Stream)	31.3	32.5
Layer 6 (Double-Stream)	**32.3**	**33.7**
Mid-to-Deep Range (Layers 6–18)		
Layer 8 (Single-Stream)	32.1	33.2
Layer 10 (Single-Stream)	31.4	32.7

Effect of sampling density. Since sampling density directly affects the representational completeness and noise level of 3D point clouds, we evaluate our method with different numbers of sampled points as inputs. The quantitative results are presented in Table 8. We observe that a moderate sampling density achieves the best performance in capturing the representational information of the foundation model. We attribute this to the balance between geometric completeness and noise accumulation. Overly sparse point clouds may fail to fully characterize the underlying 3D geometry, whereas excessively dense point clouds can introduce redundant or noisy local details that disturb feature alignment. Therefore, selecting an appropriate point density strikes an optimal balance, preserving fine-grained details while avoiding redundant noise.

Discussion on the diversity of alignment layers. The Multi-modal Diffusion Transformer is composed of Double-Stream Blocks (DSB) and Single-Stream Blocks (SSB). Notably, the DSB handles the geometric and conditional image elements independently, whereas the SSB concatenates these two modalities into a unified joint feature. To investigate which feature alignment strategy yields superior performance, we conduct alignments across different block types as well as at various depths within the same

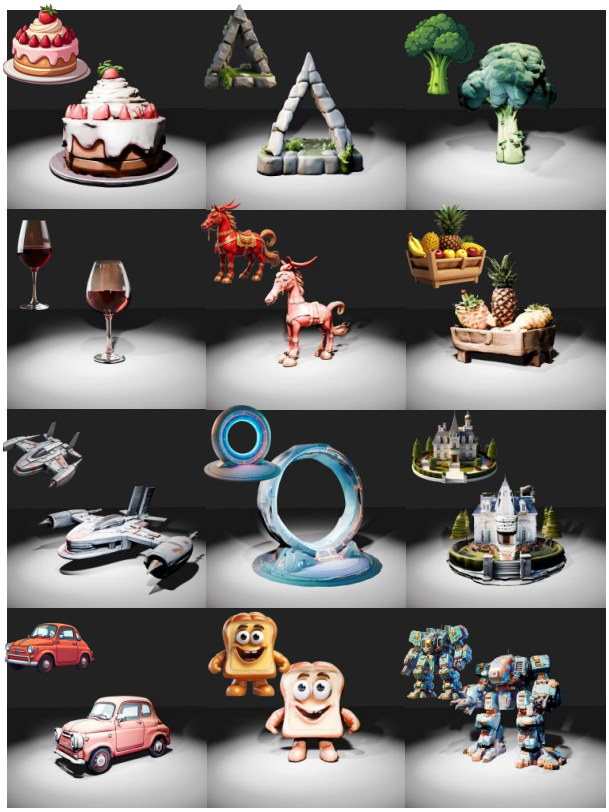


Figure 9. Visualizations across diverse asset categories demonstrate the model's robust generative capacity under various stylized conditions.

block architecture. As illustrated in Table 9, the 6th layer of the Double-Stream blocks yields superior performance. We argue that feature alignment places greater emphasis on semantic intensity. Thus, aligning at layers with stronger representations significantly influences the overall performance, with inter-modal differences having a negligible impact.

6. Conclusion

This paper presents **ROAD**, a framework designed to make 3D shape generation efficient by injecting robust priors from discriminative foundation models. We bridge the semantic gap via a reciprocal-objective strategy comprising Holistic Semantic Condensing (HSC) and Structural Optimal Alignment (SOA). This approach effectively harmonizes the unordered nature of 3D latents, ensuring precise alignment without compromising generative flexibility. Comprehensive evaluations demonstrate that ROAD achieves industrial-tier fidelity with exceptional data efficiency. We hope this work establishes a new paradigm for efficient and accessible 3D generative research.

Limitation. The generation quality is inherently bounded by the semantic capacity of the foundation model. Future work could explore integrating more advanced instructors to further push the boundaries of generation fidelity.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. of European Conference on Computer Vision*, pages 213–229, 2020.

[2] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

[3] Rui Chen, Jianfeng Zhang, Yixun Liang, Guan Luo, Weiyu Li, Jiarui Liu, Xiu Li, Xiaoxiao Long, Jiashi Feng, and Ping Tan. Dora: Sampling and benchmarking for 3d shape variational auto-encoders. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 16251–16261, 2025.

[4] Silin Cheng, Xiwu Chen, Xinwei He, Zhe Liu, and Xiang Bai. Pra-net: Point relation-aware network for 3d point cloud analysis. *IEEE Transactions on Image Processing*, 30:4436–4448, 2021.

[5] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. of Intl. Conf. on Machine Learning*, 2024.

[7] Zebin He, Mingxin Yang, Shuhui Yang, Hanxiao Sun, Xintong Han, Chunchao Guo, and Wenhan Luo. Tango3d: Towards alignment for global and local 2d-3d correspondence. *arXiv preprint arXiv:2605.19727*, 2026.

[8] Zhipeng Hu, Minda Zhao, Chaoyi Zhao, Xinyue Liang, Lincheng Li, Zeng Zhao, Changjie Fan, Xiaowei Zhou, and Xin Yu. Efficientdreamer: High-fidelity and robust 3d creation via orthogonal-view diffusion priors. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 4949–4958, 2024.

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

[10] Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. In *Proc. of Intl. Conf. on Learning Representations*, 2026.

[11] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[12] Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, et al. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025.

[13] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2025.

[14] 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 *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5307–5317, 2025.

[15] 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*, 2025.

[16] Yushi Li and George Baciu. Hsgan: Hierarchical graph learning for point cloud generation. *IEEE Transactions on Image Processing*, 30:4540–4554, 2021.

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

[18] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. In *Proc. of Advances in Neural Information Processing Systems*, pages 32653–32677, 2024.

[19] Dingkang Liang, Tianrui Feng, Xin Zhou, Yumeng Zhang, Zhikang Zou, and Xiang Bai. Parameter-efficient fine-tuning in spectral domain for point cloud learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[20] Haojia Lin, Xiawu Zheng, Lijiang Li, Fei Chao, Shanshan Wang, Yan Wang, Yonghong Tian, and Rongrong Ji. Meta architecture for point cloud analysis. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 17682–17691, 2023.

[21] Hao Liu, Hui Yuan, Junhui Hou, Raouf Hamzaoui, and Wei Gao. Pufa-gan: A frequency-aware generative adversarial network for 3d point cloud upsampling. *IEEE Transactions on Image Processing*, 31:7389–7402, 2022.

[22] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. In *Proc. of Advances in Neural Information Processing Systems*, pages 44860–44879, 2023.

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

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

[25] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision*, pages 604–621, 2022.

[26] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proc. of Intl. Conf. on Learning Representations*, 2023.

[27] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[28] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *Proc. of Intl. Conf. on Machine Learning*, pages 28223–28243, 2023.

[29] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *Proc. of European Conference on Computer Vision*, pages 214–238, 2024.

[30] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Proc. of Advances in Neural Information Processing Systems*, 35:23192–23204, 2022.

[31] Yongming Rao, Jiwen Lu, and Jie Zhou. Global-local bidirectional reasoning for unsupervised representation learning of 3d point clouds. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5376–5385, 2020.

[32] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *Proc. of Intl. Conf. on Learning Representations*, 2023.

[33] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023.

[34] Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 19615–19625, 2024.

[35] Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *Proc. of Intl. Conf. on Learning Representations*, 2026.

[36] Shuofeng Sun, Yongming Rao, Jiwen Lu, and Haibin Yan. X-3d: Explicit 3d structure modeling for point cloud recognition. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5074–5083, 2024.

[37] Yuchuan Tian, Hanting Chen, Mengyu Zheng, Yuchen Liang, Chao Xu, and Yunhe Wang. U-repa: Aligning diffusion u-nets to vits. *arXiv preprint arXiv:2503.18414*, 2025.

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

[39] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. In *Proc. of Advances in Neural Information Processing Systems*, 2022.

[40] Ziyu Wan, Despoina Paschalidou, Ian Huang, Hongyu Liu, Bokui Shen, Xiaoyu Xiang, Jing Liao, and Leonidas Guibas. Cad: Photorealistic 3d generation via adversarial distillation. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 10194–10207, 2024.

[41] Changshuo Wang, Meiqing Wu, Siew-Kei Lam, Xin Ning, Shangshu Yu, Ruiping Wang, Weijun Li, and Thambipillai Srikanthan. Gpsformer: A global perception and local structure fitting-based transformer for point cloud understanding. In *Proc. of European Conference on Computer Vision*, pages 75–92. Springer, 2024.

[42] Chuxin Wang, Yixin Zha, Jianfeng He, Wenfei Yang, and Tianzhu Zhang. Rethinking masked representation learning for 3d point cloud understanding. *IEEE Transactions on Image Processing*, 34:247–262, 2024.

[43] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023.

[44] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5294–5306, 2025.

[45] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Proc. of Advances in Neural Information Processing Systems*, 2023.

[46] Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. In *Proc. of Advances in Neural Information Processing Systems*, 2025.

[47] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

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

[49] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *Proc. of Advances in Neural Information Processing Systems*, pages 33330–33342, 2022.

[50] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024.

[51] 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 *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 21469–21480, 2025.

[52] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 1179–1189, 2023.

[53] 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 *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 27091–27101, 2024.

[54] Xianghui Yang, Guosheng Lin, and Luping Zhou. Single-view 3d mesh reconstruction for seen and unseen categories. *IEEE Transactions on Image Processing*, 32:3746–3758, 2023.

[55] 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 *Proc. of IEEE Intl. Conf. on Computer Vision*, 2025.

[56] Haochen Yu, Weixi Gong, Jiansheng Chen, and Huimin Ma. Homugan: A 3d-aware gan with the method of cylindrical spatial-constrained sampling. *IEEE Transactions on Image Processing*, 34:320–334, 2024.

[57] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *Proc. of Intl. Conf. on Learning Representations*, 2025.

[58] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022.

[59] Songyang Zhang, Shuguang Cui, and Zhi Ding. Hypergraph spectral analysis and processing in 3d point cloud. *IEEE Transactions on Image Processing*, 30:1193–1206, 2020.

[60] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Vide-

orepa: Learning physics for video generation through relational alignment with foundation models. In *Proc. of Advances in Neural Information Processing Systems*, 2025.

[61] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 16259–16268, 2021.

[62] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *Proc. of Intl. Conf. on Learning Representations*, pages 46766–46782, 2024.

[63] Yan Zhou, Dewang Ye, Huaidong Zhang, Xuemiao Xu, Huajie Sun, Yewen Xu, Xiangyu Liu, and Yuexia Zhou. Recurrent diffusion for 3d point cloud generation from a single image. *IEEE Transactions on Image Processing*, 2025.