Know3D: Prompting 3D Generation with Knowledge from Vision-Language Models

Wenyue Chen[1,2*], Wenjue Chen[4*], Peng Li[3], Qinghe Wang[5], Xu Jia[5], Heliang Zheng[2], Rongfei Jia[2], Yuan Liu[3†], and Ronggang Wang[1†]

[1] Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, Shenzhen Graduate School, Peking University, Shenzhen, China
[2] Math Magic, Beijing, China
[3] The Hong Kong University of Science and Technology, Hong Kong, China
[4] Peking University, Beijing, China
[5] Dalian University of Technology, Dalian, China

Abstract. Recent advancements in 3D generation have significantly enhanced the fidelity and geometric details of synthesized 3D assets. However, due to the inherent ambiguity of single-view input and the lack of robust global structural priors–caused by limited 3D training data–the unseen regions generated by existing models remain stochastic and difficult to control. This often results in geometries that are either physically implausible or misaligned with user intent. In this paper, we propose Know3D, a novel framework designed to incorporate rich knowledge from Multimodal Large Language Models (MLLMs) into 3D generation processes. By leveraging latent hidden-state injection, Know3D supports language-controllable generation of the back-view for 3D assets. We utilize a VLM-diffusion-based architecture: the Vision Language Model (VLM) is used to provide high-level semantic understanding, while the diffusion model serves as a bridge, transferring semantic knowledge into the 3D generation model. Extensive experiments demonstrate that Know3D effectively bridges the gap between abstract textual instructions and the geometric reconstruction of invisible regions. By transforming the traditionally stochastic back-view hallucination into a semantically controllable process, Know3D offers a promising direction for highly plausible and user-friendly 3D generation in the future. Project page: `https://xishuxishu.github.io/Know3D.github.io/`.

Keywords: Image-to-3D Generation · Multimodal Foundation Models · Semantic Control

1 Introduction

High-quality 3D assets are essential to modern workflows across gaming, film, and embodied AI. Because manual modeling remains prohibitively labor-intensive,

* Equal contribution.
† Corresponding authors.

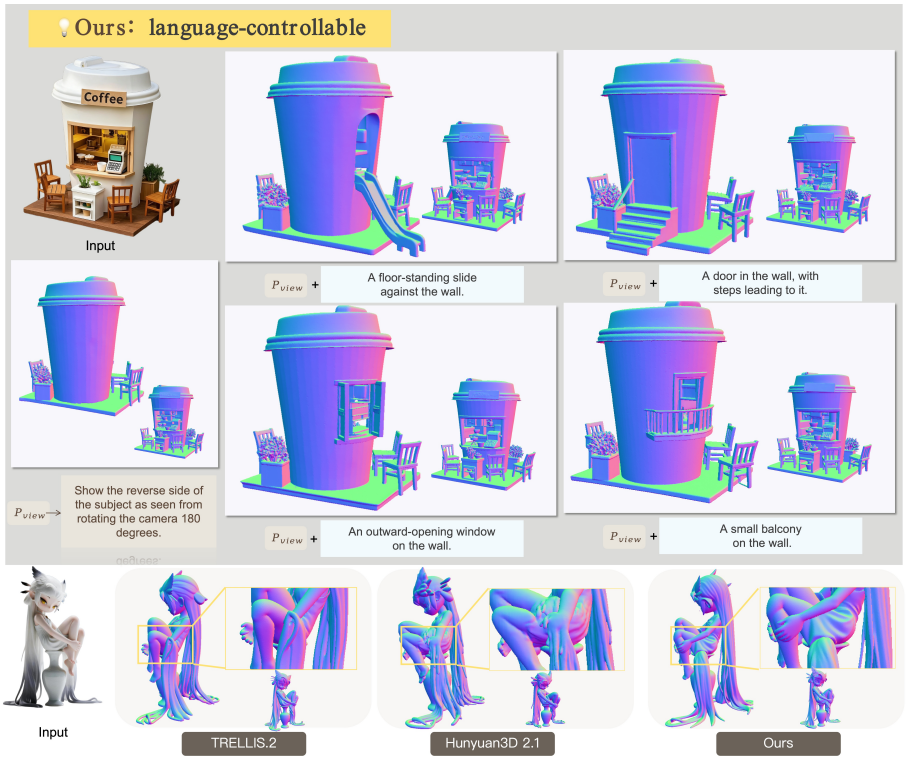

Fig. 1: We propose Know3D, a knowledge-guided 3D generation framework that enables semantic control over the back-view of target objects. As shown in the coffee-cup scene, by simply varying the text, Know3D can synthesize diverse and structurally consistent back-view components. The bottom-row results demonstrate the potential of our method to improve the structural plausibility of unseen parts in 3D generation.

automating 3D asset generation has become a critical challenge for the vision and graphics communities. Recently, 3D generative modeling [5, 8, 10, 17, 20, 24, 30, 57, 59, 60, 67] has advanced at a rapid pace. These breakthroughs have significantly enhanced both the fine-grained geometric details and visual fidelity of 3D assets.

Despite these advancements, generating 3D assets from a single image remains a fundamentally ill-posed problem due to the inherent ambiguity of single-view observations. Image-to-3D generative models [20, 59, 60] learn to map 2D observations to 3D shape by modeling the distribution of their training data. Since the input image only contains the information of the visible part, the synthesis of invisible parts relies on the model's internalized priors. While existing models are capable of hallucinating the occluded back-view, the synthesis remains predominantly stochastic and uncontrollable, and prone to two critical failure modes: (1) producing outputs that deviate from the user's creative or semantic intentions, as existing models inherently lack the ability to align the unob-

served region synthesis with user-specified semantic constraints; (2) generating geometrically implausible structures that violate basic semantic commonsense constraints, as shown in Fig. 1. These failure modes can be largely attributed to data constraints. Compared to the internet-scale abundance of images, text, and videos, 3D datasets [11, 12, 21, 68] are relatively limited in both quantity and diversity. Consequently, the world knowledge and structural common sense internalized by 3D generation models are naturally constrained. Further, high-quality, semantically aligned text-3D paired data remains less abundant.

Bridging this gap requires going beyond visual priors by incorporating additional knowledge, enabling models to infer unobserved structures from visible evidence. In particular, modern vision-language models (VLMs) have already learned rich semantic knowledge and commonsense reasoning from internet-scale multimodal data. If such knowledge can be effectively transferred to 3D generation models, it may provide valuable guidance for inferring unobserved object structures.

However, effectively prompting VLM knowledge into 3D generation presents its own challenges. A naive approach would be to directly use LLMs or VLMs to generate 3D representations in an autoregressive manner, as explored by recent shape LLM works [14, 54, 65]. Yet, these methods have thus far underperformed compared to dedicated 3D generative models. Moreover, this autoregressive paradigm alters the model's pretrained knowledge priors. Forcing them into a constrained 3D generation task disrupts their original semantic capabilities.

Another possible strategy is to directly feed VLM representations into 3D generation networks. However, such representations are typically highly abstract and lack explicit geometric grounding. As a result, they do not align well with the spatially structured feature spaces required for accurate 3D shape synthesis.

To address this challenge, we propose Know3D, a novel framework that prompts 3D generation with knowledge by leveraging the rich semantic understanding and commonsense reasoning capabilities of vision-language models (VLMs), thereby achieving enhanced controllability and better plausibility in 3D generation. Instead of directly injecting abstract VLM representations, we leverage a multimodal diffusion model as an intermediate bridge that translates semantic knowledge into image-space structural priors.

Specifically, we utilize a VLM-diffusion-based model (Qwen-Image-Edit [56]), where the VLM [2] is responsible for semantic understanding and provides guidance for image generation, and the diffusion model is responsible for generating images of the unobserved parts based on this guidance. The image-space structural priors serve as a medium to provide both semantic and structural information, thereby enabling semantically controllable 3D generation.

Although the Qwen-Image-Edit model demonstrates strong semantic understanding and can generate novel views based on prompts, it has two notable shortcomings: first, it often misinterprets spatial orientation, such as failing to accurately generate a "back-view" of an object; second, it frequently alters the subject's original pose or action in the output. Thus, we fine-tune Qwen-Image-Edit to improve the spatial awareness for better stability. Note that we annotate

the corresponding textual description of the back-view for training to enable the generation control of the back-view.

We explored how to use the image-space structural priors as an intermediate medium to prompt the knowledge from VLMs into 3D generation models, experimenting with different designs for connecting them. Specifically, we experimented with (1) directly using the fully denoised VAE latent from MMDiT, (2) decoding this fully denoised VAE latent into an image and then extracting features via DINOv3 [41], and (3) directly using the hidden states from the intermediate layers of MMDiT during the denoising process. Among these, directly using the hidden states from the intermediate layers of MMDiT demonstrates better spatial and semantic awareness and consequently achieves the best overall performance.

Evaluations on HY3D-Bench [21] show that Know3D achieves competitive performance against state-of-the-art single-view 3D generation methods in semantic consistency with the conditional image. Moreover, our framework enables language-controllable generation of unseen backside regions, as shown in Fig. 1.

2 Related Works

2.1 Native Single-view 3D Generation

In recent years, native single-view 3D generation based on diffusion models has entered a period of rapid evolution, driven primarily by the advancement of 3D latent representations which have converged into two dominant paradigms: the Vector Set (VecSet) [20, 23, 25, 27–29, 66, 67, 70] approach that prioritizes global perception and high compression rates, and the Sparse Voxel approach [17, 30, 38, 57, 59, 60, 64] that excels in local control and complex topological expression. Recent works have begun exploring the complementary fusion of these two paradigms. For instance, some approaches employ decoupled "coarse-to-fine" refinement frameworks to resolve the conflict between global structure and local geometry [8,22,24], LATTICE [24] introduces semi-structured hybrid representations that inject spatial anchors into latent sets to enhance detail fidelity. Driven by these outstanding works, current models have achieved significant breakthroughs in both geometric and appearance fidelity. However, existing single-view 3D generation methods still have limitations in generating unobserved regions. This is mainly due to the limited information from a single view and the constraints of 3D training data.

2.2 Text-to-3D Generation

As a groundbreaking work in text-to-3D generation, DreamFusion [36] pioneers score distillation from pre-trained 2D diffusion models to optimize 3D assets, with numerous subsequent works [32, 33, 36, 44, 45, 55] further refining this distillation pipeline for better generation performance. Native text-to-3D methods enable end-to-end generation directly in 3D representation spaces, with remarkable

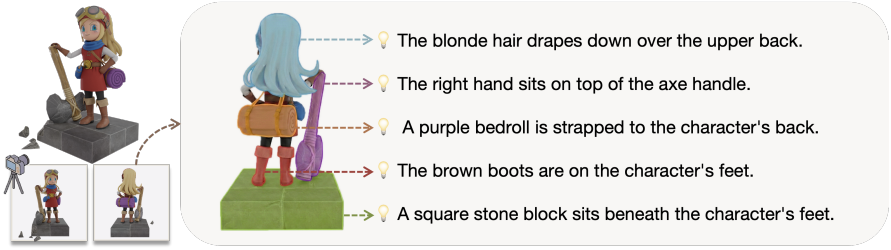

Fig. 2: Text annotation pipeline. Given paired front and back views of a 3D asset, a VLM generates initial part-level descriptions.

progress achieved in recent works [27–29,57,60,69,70]. Nevertheless, the controllability and fine-grained geometric accuracy of such paradigms still lag behind those of image-guided image-to-3D approaches. Some works [7,9,14,37,40,54,65] explored Multimodal large language models for 3D generation, yet these methods are constrained by limited representation resolution, failing to achieve high-fidelity 3D content generation.

2.3 Unified Multimodal Models

Recent research has focused on unified multimodal models for joint image understanding and generation. Existing methods fall into these paradigms: unified autoregressive models [6, 15, 16, 42, 47, 50, 53, 58], unified diffusion models [31,39,43,52,63], decoupled LLM-diffusion frameworks [3,4,35,56], and hybrid AR-diffusion architectures [13,61,71]. While the 3D generation field has also been progressing rapidly in recent years, its overall development timeline lags behind that of the multimodal domain. Achieving semantic control in 3D generation remains challenging due to the limitation of 3D training data, which restricts the scale of multimodal pre-training compared to 2D domains. Furthermore, while text descriptions are highly abstract and lack geometric constraints, 3D generation requires explicit spatial, textural, and structural priors. To bridge this gap, we explore leveraging multimodal diffusion models as an intermediate bridge between vision-language knowledge and 3D generation. Instead of directly injecting abstract VLM representations, we observe that the intermediate hidden states of diffusion transformers encode rich spatial and structural information during the denoising process.

3 Method

Overview. Given a single input image I and a textual description T of the target object's back-view, our goal is to synthesize a complete 3D representation $\mathcal{V}$. To provide semantic cues for the unseen side, we fine-tune Qwen-Image-Edit on paired front-back view data to generate a plausible back-view image conditioned on both I and T (Section 3.1). Then, we propose Know3D (Section 3.2), a

knowledge-guided 3D generation framework that enables semantic control over the back-view of target objects. The overall pipeline is illustrated in Fig. 3.

3.1 Semantic-Aware Front-Back View Generation

In this section, we aim to fine-tune Qwen-Image-Edit-2511 [56] to generate reasonable back-view images from a given image. Even though the Qwen-Image-Edit model already shows a strong semantic understanding of the input image and can generate novel view images following the prompt. It still lacks a strong spatial awareness to understand the "back-view" and often generates images from incorrect viewpoints. In addition to viewpoint inaccuracies, it also tends to alter the subject's original pose during generation. Thus, we fine-tune Qwen-Image-Edit to improve the spatial awareness for better stability. Note that we annotate the corresponding textual description of the back-view for training to enable the generation control of the back-view.

Dataset Construction We construct the training data from high-quality 3D assets. For each asset, we render $2N$ images using uniform azimuth sampling with random elevation. We select N views as front views and pair each with its opposite view to form front–back pairs $(I_{\text{front}}, I_{\text{back}})$. To enable semantic control for back-view generation, we annotate textual descriptions for each front–back image pair. For each front–back pair $(I_{\text{front}}, I_{\text{back}})$, we annotate a set of textual descriptions for the salient components visible in the back-view. The output is a description set $D = \{d_1, d_2, \ldots, d_m\}$, where each d_i describes one back-view component, as shown in Fig. 2.

Training Strategy and Objective. To achieve stable back-view generation with text prompt control, we design a stochastic prompting strategy and optimize the model using the Conditional Flow Matching (CFM) objective [34, 49].

Stochastic Prompt Construction. To enable controllable generation, we construct the conditioning prompt:

$$P = P_{\text{view}} \oplus P_{\text{back}}, \tag{1}$$

P_{view} is a fixed prompt describing the 180° camera rotation, while P_{back} is randomly sampled from the component-level description set D with probability 0.5. This stochastic prompting strategy enables the model to learn both unconditional back-view generation and semantically controlled generation.

Training Objective. Following Qwen-Image-Edit [56], we extract multimodal hidden states $H_{\text{vlm}} = \text{Qwen2.5-VL}(I_{\text{front}}, P)$ from the vision-language model [2], and obtain the spatial latent condition $Z_{\text{front}} = \mathcal{E}(I_{\text{front}})$ via a VAE encoder [51] $\mathcal{E}$. Let $Z_{back} = \mathcal{E}(I_{back})$ denote the target latent of the back-view. Given a noisy latent $Z_t = (1-t)Z_{back} + t\epsilon$ at timestep $t \in [0, 1]$, the vector field estimator

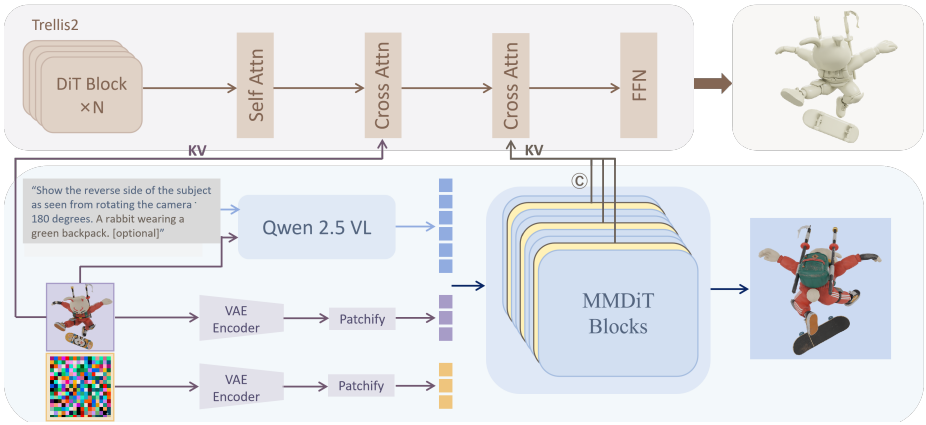

Fig. 3: Overview. Given an input image and a text prompt, Know3D leverages Qwen-Image-Edit-2511 [56], which integrates Qwen2.5-VL with a Diffusion Transformer (DiT). Qwen2.5-VL [2] encodes the multimodal inputs to guide a latent denoising process, enabling semantically aligned backside generation. The latent hidden states from intermediate DiT layers, which embed rich semantic and structural priors, are subsequently injected as conditioning signals into the 3D generation model to guide complete shape synthesis.

v_ϕ is optimized to predict the velocity field via the Conditional Flow Matching objective [34, 49]:

$$\mathcal{L}_{\text{CFM}}(\phi) = \mathbb{E}_{t, Z_{\text{back}}, \epsilon} \| v_\phi(Z_t, t, H_{\text{vlm}}, Z_{\text{front}}) - (\epsilon - Z_{\text{back}}) \|_2^2. \tag{2}$$

3.2 Prompting 3D Generation with VLMs

In this section, we introduce how to use image features as an intermediate medium to prompt knowledge from vision-language models (VLMs) for 3D generation models. The more straightforward approach is to directly inject the images generated by Qwen-Image. However, this process involves VAE decoding and the re-extraction of features by DINOv3 [41], making the workflow relatively cumbersome. Moreover, it relies on high-precision pixel-level restoration. If the quality of the generated images is insufficient, erroneous results will directly impact the 3D generation process. An ideal feature should possess the following characteristics: (1) sufficient spatial awareness to facilitate learning by 3D generation models, and (2) a certain degree of semantic awareness and robustness. We found that the hidden states of the intermediate layers of MMDiT inherently possess strong spatial awareness and rich semantic information [19, 26], enabling them to better guide 3D generation.

Knowledge Extraction and Prompting. Qwen2.5-VL [2] encodes the input front-view image and text prompt into high-level semantic features, while the

VAE encoder extracts visual features from the front-view input. These representations guide the MMDiT through the full iterative denoising process. We extract intermediate latent hidden states from n MMDiT layers at a specific denoising timestep t [1,19,46], and concatenate these layer-wise features to form the structural-semantic conditioning signal $H_{\mathrm{DiT}} = \mathrm{Concat}(h^{(1)}, h^{(2)}, \ldots, h^{(n)})$.

Building upon TRELLIS2 [59], we design a parallel cross-attention branch for H_{DiT} injection. We retain the backbone's original self-attention and image-conditioned cross-attention layers intact to avoid interference with pre-trained 3D generation priors. H_{DiT} is first linearly projected and then layer-normalized to obtain the projected feature H'_{DiT}, which serves as keys and values for the new cross-attention layer. Its output is scaled by a zero-initialized linear layer for stable training. The modified residual fusion process is formulated as:

$$\begin{aligned} F_{\mathrm{sa}} &= \text{Self-Attn}(F), \\ \Delta F_{\mathrm{img}} &= \text{Cross-Attn}(F_{\mathrm{sa}}, F_{\mathrm{img}}, F_{\mathrm{img}}), \\ \Delta F_{\mathrm{dit}} &= \text{ZeroLinear}(\text{Cross-Attn}(F_{\mathrm{sa}}, H'_{\mathrm{DiT}}, H'_{\mathrm{DiT}})), \\ F_{\mathrm{out}} &= F + \Delta F_{\mathrm{img}} + \Delta F_{\mathrm{dit}}. \end{aligned} \tag{3}$$

3D Geometry Generation With the structural-semantic signal H_{DiT} and front-view feature F_{front} as dual conditioning signals, our 3D generation follows the two-stage paradigm of TRELLIS2 [59]:

$$\begin{aligned} V_{\mathrm{ss}} &= D_{\mathrm{ss}}\left(G_{\mathrm{ss}}\left(z, \tau, F_{\mathrm{front}}, H_{\mathrm{DiT}}\right)\right), \\ V_{\mathrm{geo}} &= D_{\mathrm{geo}}\left(G_{\mathrm{geo}}\left(z, \tau, F_{\mathrm{img}}, H_{\mathrm{DiT}} \mid V_{\mathrm{ss}}\right)\right), \end{aligned} \tag{4}$$

where z is standard Gaussian noise, τ is the diffusion timestep. The first stage generates a coarse sparse structure V_{ss} to model the global topological prior, and the second stage recovers high-fidelity fine geometry V_{geo} conditioned on V_{ss}.

Training Objective. Both stages are optimized with the Conditional Flow Matching (CFM) objective [34, 49],

$$\mathcal{L}_{3D} = \mathbb{E}_{\tau, x_0, \epsilon} \| v_\theta(x_\tau, \tau, F_{\mathrm{img}}, H_{\mathrm{DiT}}) - (\epsilon - x_0) \|_2^2, \tag{5}$$

where x_0 is the ground-truth 3D geometric latent, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian noise.

4 Experiments

4.1 Experiment Setup

Dataset. For Semantic-Aware Front-Back View Generation Training, we use 5k high-quality 3D assets selected from the TexVerse dataset [68]. For each asset, the field of view (FoV) is sampled from $\{35°, 50°, 85°, 105°, 135°\}$ and elevation from $[-15°, 45°]$, with both fixed per asset but randomized across assets.

We render 12 uniformly spaced azimuthal views over 360° per mesh, forming 6 front-back pairs, and annotate all of them. For 3D generation training, we use 60k meshes from TexVerse. For each mesh, we render two sets of views: a perturbation-free set and a perturbed set. In the perturbation-free set, azimuth views are spaced every 45° and grouped into four front–back pairs. In the perturbed set, we apply random FoV scaling and small angular offsets to simulate viewpoint perturbations, forming four perturbed pairs. For evaluation, we conduct quantitative analysis and comparisons with baselines on the HY3D-Bench [21] dataset. For the ablation study, we randomly selected a subset of 100 3D assets in TexVerse [68] dataset that not in our training data.

Training Details. In Semantic-Aware Front-Back View Generation, we adopt Qwen-Image-Edit-2511 [56] as our foundational pre-trained model, and fine-tune it using Low-Rank Adaptation(LoRA) [18]. All experiments are conducted on 32 NVIDIA A800 GPUs with a global batch size of 32. We set the rank of the LoRA adapter to 64 for all trainable attention layers of the backbone model. For 3D generation, we freeze the parameters of the pre-trained Qwen-Image-Edit-2511 [56] to preserve its generalizable visual priors. For the original Trellis2 network, we apply LoRA [18] fine-tuning with a rank of 64. In contrast, the newly added condition layers, which are designed to integrate semantic-aware front-back view signals, are fully fine-tuned to ensure effective modulation of the 3D generation process. This stage is trained on 32 NVIDIA A800 GPUs with a global batch size of 64.

Metrics. To compare with baseline, we use ULIP [62] and Uni3D [72] to measure the semantic consistency between images and generated meshes. For the ablation study, we only trained the first stage (sparse voxel generation). Therefore, we evaluate the performance using IoU and Chamfer Distance (CD).

4.2 Generation Controllability and Quality

In this section, we present a comparison of Know3D with current state-of-the-art methods, as well as a qualitative analysis of its semantic controllability.

Comparison with Baselines. We evaluate the generation quality of Know3D by comparing with single- and multi-view 3D generation methods.

Comparison with Single-Image-to-3D Generation Methods. We conduct comparative experiments with several state-of-the-art, open-source single-image-to-3D generation methods, including Hunyuan3D-2.1 [20], TRELLIS.2 [59], TRELLIS [60], Step1X-3D [28], Hi3DGen [64], and Direct3D-S2 [57]. As shown in Tab. 1, Know3D achieves competitive ULIP and Uni3D scores, indicating effective semantic alignment between the generated meshes and input images. Fig. 6 presents the qualitative comparison of back-view geometry between Know3D

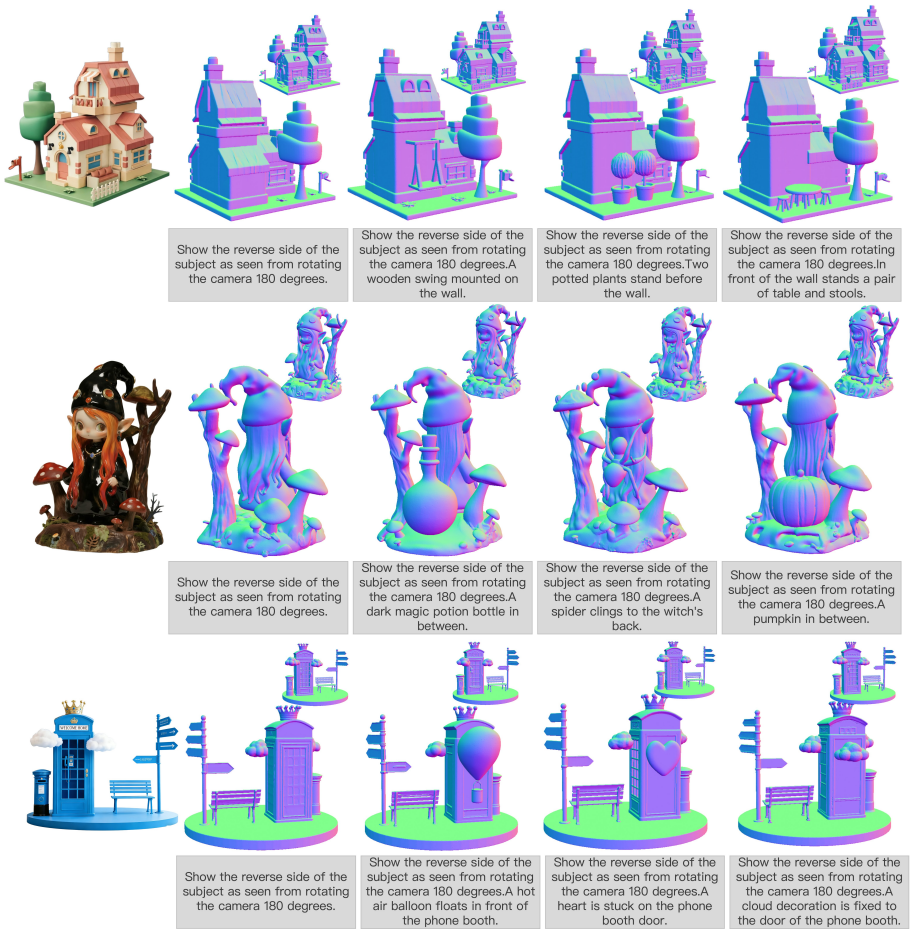

Fig. 4: Back-View Semantic Control Visualization. The leftmost column shows the input front-view condition images. The remaining columns present back-view normal renderings of meshes generated under different text prompts, demonstrating flexible semantic control over back-side components.

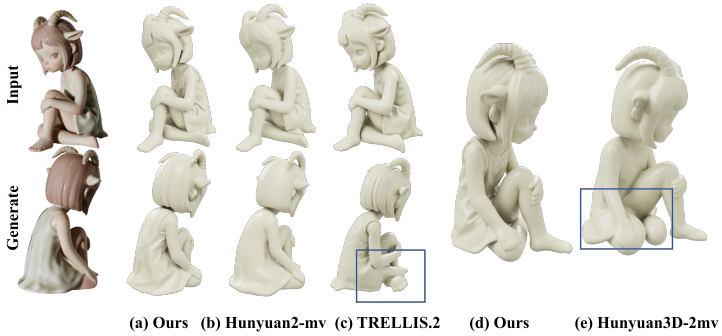

Fig. 5: Visual Comparison between Know3D and Hunyuan3D-2mv. As shown in (a) and (b), both Know3D and Hunyuan3D-2mv [48] achieve reasonably good alignment between the input and generated viewpoints. However, for novel viewpoints, Hunyuan3D-2mv produces implausible structures(d), whereas Know3D still yields reasonable results(e).

and SOTA baselines, visualized via surface normal maps. Our method leverages semantic knowledge to harness the intrinsic understanding of object attributes within pre-trained multimodal models, demonstrating its potential to enhance the structural plausibility of unseen components in 3D generation.

Comparison with Hunyuan3D-2mv. In addition, to analyze the effect of directly using synthesized views as multi-view inputs, we further construct a simple baseline by feeding the input front view together with the generated back-view into a multi-view 3D generation model. In this setting, we adopt Hunyuan3D-2mv [48] as the multi-view baseline for comparison. Despite being trained on large-scale 3D data for 3D generation, Hunyuan3D-2mv is outperformed by Know3D in terms of ULIP and Uni3D scores as shown in Tab. 1. This validates the effectiveness of our Knowledge Extraction and Prompting design. As illustrated in Fig. 5, although both methods achieve reasonable alignment with the conditioned views (a),(b), Hunyuan3D-2mv generates distorted and implausible geometries (d), whereas Know3D maintains consistent structures (e).

Back-View Semantic Controllability. Know3D's key advantage over existing methods is its ability to semantically control back-view content via natural language instructions. Existing single-view 3D generation methods can only replicate the visible front view, while back-view generation remains stochastic and cannot respond to user-specified requirements for occluded regions. Fig. 4 demonstrates our model's flexible semantic controllability over the back-view. The leftmost column shows the input front-view images. The second column presents the unconstrained back-view completion results of Know3D, which already produce geometrically plausible and semantically consistent structures aligned with the front view. The subsequent columns present semantically con-

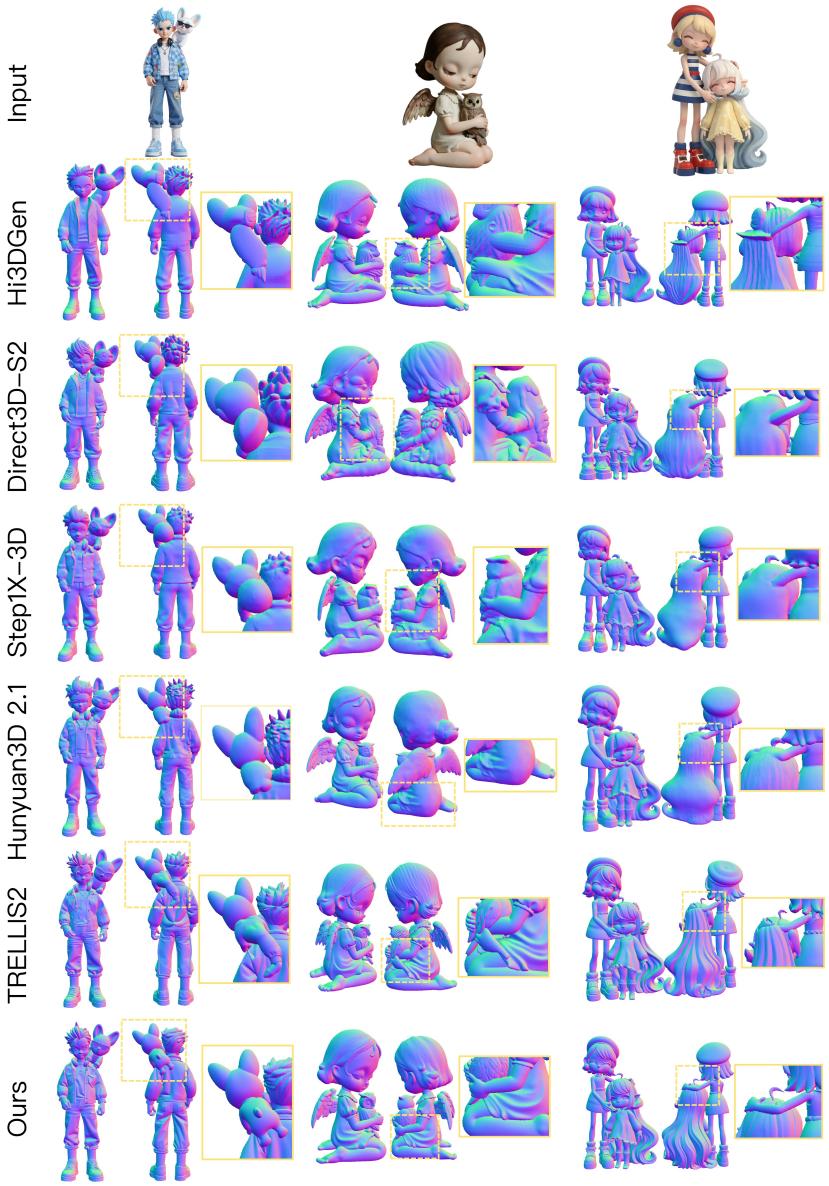

Fig. 6: Visual Comparison between Know3D and Current SOTA Baselines. By leveraging semantic knowledge, our method taps into the intrinsic understanding of object attributes within pre-trained multimodal models, showing potential for improving the structural plausibility of unseen parts in 3D generation.

Table 1: Quantitative comparison on HY3D-Bench [21]. The best results are in bold and the second-best are underlined. ↑ indicates higher is better.

	HY3D-Bench-test		HY3D-Bench-val	
Model	ULIP ↑	Uni3D ↑	ULIP ↑	Uni3D ↑
Direct3D-S2 [57]	0.2034	0.3292	0.2040	0.3303
Hi3DGen [64]	0.2082	0.3206	0.2030	0.3213
Step1X-3D [28]	0.2068	0.3233	0.2033	0.3265
TRELLIS [60]	<u>0.2143</u>	0.3421	**0.2133**	<u>0.3464</u>
TRELLIS.2 [59]	0.1948	0.3308	0.1967	0.3358
Hunyuan3D-2MV [48]	0.2108	0.3413	0.2111	0.3400
Hunyuan3D-2.1 [20]	0.2140	<u>0.3434</u>	0.2122	0.3433
Ours	**0.2174**	**0.3518**	<u>0.2127</u>	**0.3512**

trolled back-view synthesis results guided by different text prompts. These results show that our model can modify the unseen back-side content according to user instructions, while preserving geometric consistency with the original front view.

4.3 Ablation Study

We conduct ablation studies to explore two questions in our knowledge prompting design. First, we investigate how the choice of MMDiT timestep for feature extraction influences generation performance—do earlier or later stages of the denoising process provide more useful priors. Second, we compare different feature representations—MMDiT [56] latent hidden states, VAE encoder [51] features, and DINOv3 [41] features—to assess what type of information is most effective for guiding 3D generation. We rendered front and back views of a 100-randomly-selected 3D assets subset and extracted their sparse voxels as ground truth for subsequent comparisons. We directly use the ground truth front and back views as the condition to compare the capabilities of different features. For all ablation experiments, we adhered to the same experimental setup: training the first-stage DiT, which generates sparse voxels on 60k training samples with a batch size of 64 for 70k steps. For the original parameters of TRELLIS.2, we employed LoRA (rank=64) for adaptation, while the newly added modules were fully fine-tuned.

Different Timesteps in MMDiT. We evaluate the impact of the denoising timestep t used to extract the MMDiT hidden states, considering $t \in \{0, 0.25, 0.5, 0.75\}$. As shown in Table 2, $t = 0.25$ achieves the best overall performance, yielding the highest IoU and lowest Chamfer Distance (CD). We hypothesize that at this intermediate stage ($t = 0.25$), the MMDiT has already resolved the global layout and core semantic components of the back-view. Features from earlier timesteps may focus on low-level pixel-wise details that are less aligned with 3D structural priors, whereas later stages are subject to increased noise interference, leading to less reliable guidance for the 3D generation backbone.

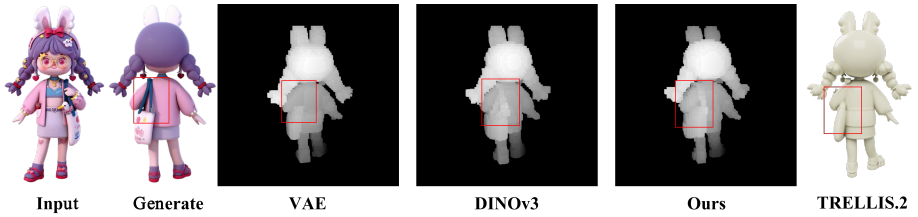

Fig. 7: Visualization of Different Feature Representations to Prompt 3D Generation. When Qwen-Image generates an unreasonable back-view, using different feature representations to prompt 3D generation will lead to different results. The hidden states from the intermediate layers of MMDiT during the denoising process still generate reasonable results, while the VAE fails to inject the back-view information, making it quite similar to the unreasonable results of the original TRELLIS.2. Additionally, when the generated image is directly passed through DINOv3, DINOv3 tends to fit the erroneous results.

Table 2: Ablation on MMDiT timestep. The best results are in bold and the second-best are underlined. ↑/↓ indicate higher/lower is better.

Timestep t	IoU ↑	CD ↓
0.0	0.343	2.376
0.25	**0.352**	**2.262**
0.5	<u>0.349</u>	<u>2.272</u>
0.75	0.336	2.452

Table 3: Ablation on different feature representations. The best results are in bold and the second-best are underlined. ↑/↓ indicate higher/lower is better.

Feature Source	IoU ↑	CD ↓
VAE encoder	0.308	2.803
DINOv3	<u>0.342</u>	<u>2.385</u>
MMDiT ($t = 0.25$)	**0.352**	**2.262**

Different Feature Representations to Prompt. We further investigate the effectiveness of various feature representations for guiding 3D generation. Specifically, we experiment with (1) directly using the fully denoised VAE [51] latent from MMDiT, (2) decoding this fully denoised VAE latent into an image and then extracting features via DINOv3 [41], and (3) directly using the hidden states from the intermediate layers of MMDiT during the denoising process.

As shown in Table 3, the MMDiT hidden states extracted at timestep $t = 0.25$ consistently outperform other feature representations on both metrics, achieving the best overall performance. This suggests that MMDiT hidden states contain rich priors of 3D semantics and structure. We have also observed similar phenomena in recent studies [19]. In contrast, the VAE encoder features yield the lowest performance. This may be because VAE latents are primarily optimized for pixel-level reconstruction, which tends to preserve low-level appearance information while discarding higher-level semantic and structural cues that are important for 3D generation. Moreover, as shown in Fig. 7, when leveraging the pipeline of Qwen-Image to guide 3D generation, inconsistencies may occasionally arise. For instance, the front view might depict a single-shoulder bag, while

the generated back-view incorrectly includes two straps. Using different feature representations to prompt 3D generation will lead to different results. The VAE fails to inject the back-view information, making it quite similar to the impossible results of the original TRELLIS.2. Additionally, when the generated image is directly passed through DINOv3, DINOv3 tends to fit the erroneous results. The hidden states from the intermediate layers of MMDiT during the denoising process still generate reasonable results. This shows that it not only has strong 3D perception ability but also contains rich semantic information.

5 Limitation and Conclusion

In this paper, we propose Know3D, a novel knowledge-guided 3D generation framework designed to address the inherent ambiguity and lack of semantic control in single-view 3D synthesis. By leveraging a large multimodal model as a "semantic brain", we successfully bridge the gap between abstract textual instructions and the geometric reconstruction of the unobserved regions, transforming the traditionally stochastic back-view hallucination into a semantically-controllable process. We extract rich structural–semantic priors from the intermediate MMDiT layers to prompt the 3D generation model. Our ablation studies further confirm that intermediate denoising states from the multimodal diffusion process capture essential 3D structural and semantic priors.

Limitation. While Know3D achieves semantic controllability in 3D generation by leveraging multimodal priors, the structural robustness of the generated assets is still influenced by the underlying multimodal foundation models. When the multimodal foundation model cannot fully understand the instructions, then the 3D generation will still be misled to incorrect 3D shapes. This could potentially be alleviated by adopting stronger MLLMs or exploring more effective ways of leveraging multimodal guidance and information injection in the 3D generation process.

Acknowledgements

This work was financially supported by the Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology (Grant No. 2024B1212010006), the Outstanding Talents Training Fund in Shenzhen, the Shenzhen Science and Technology Program (Grant Nos. SYSPG20241211173440004 and KJZD20230923114707016), the R24115SG MIGU-PKU META VISION TECHNOLOGY INNOVATION LAB., and the National Natural Science Foundation of China (Grant No. 62272142).

References

1. Baade, A., Chan, E.R., Sargent, K., Chen, C., Johnson, J., Adeli, E., Fei-Fei, L.: Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. arXiv preprint arXiv:2602.11401 (2026)

2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
4. Chen, J., Xue, L., Xu, Z., Pan, X., Yang, S., Qin, C., Yan, A., Zhou, H., Chen, Z., Huang, L., et al.: Blip3o-next: Next frontier of native image generation. arXiv preprint arXiv:2510.15857 (2025)
5. Chen, R., Zhang, J., Liang, Y., Luo, G., Li, W., Liu, J., Li, X., Long, X., Feng, J., Tan, P.: Dora: Sampling and benchmarking for 3d shape variational auto-encoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16251–16261 (2025)
6. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
7. Chen, Y., He, T., Huang, D., Ye, W., Chen, S., Tang, J., Chen, X., Cai, Z., Yang, L., Yu, G., et al.: Meshanything: Artist-created mesh generation with autoregressive transformers. arXiv preprint arXiv:2406.10163 (2024)
8. 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)
9. Chen, Y., Lan, Y., Zhou, S., Wang, T., Pan, X.: Sar3d: Autoregressive 3d object generation and understanding via multi-scale 3d vqvae. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28371–28382 (2025)
10. Chen, Z., Tang, J., Dong, Y., Cao, Z., Hong, F., Lan, Y., Wang, T., Xie, H., Wu, T., Saito, S., et al.: 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26576–26586 (2025)
11. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023)
12. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
13. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
14. Fang, S., Shen, I., Wang, Y., Tsai, Y.H., Yang, Y., Zhou, S., Ding, W., Igarashi, T., Yang, M.H., et al.: Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14061–14072 (2025)
15. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024)

16. Geng, Z., Wang, Y., Ma, Y., Li, C., Rao, Y., Gu, S., Zhong, Z., Lu, Q., Hu, H., Zhang, X., et al.: X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. arXiv preprint arXiv:2507.22058 (2025)
17. He, X., Zou, Z.X., Chen, C.H., Guo, Y.C., Liang, D., Yuan, C., Ouyang, W., Cao, Y.P., Li, Y.: Sparseflex: High-resolution and arbitrary-topology 3d shape modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14822–14833 (2025)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
19. Huang, Z., Li, X., Lv, Z., Rehg, J.M.: How much 3d do video foundation models encode? arXiv preprint arXiv:2512.19949 (2025)
20. Hunyuan3D, T., Yang, S., Yang, M., Feng, Y., Huang, X., Zhang, S., He, Z., Luo, D., Liu, H., Zhao, Y., et al.: Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. arXiv preprint arXiv:2506.15442 (2025)
21. Hunyuan3D, T., Zhang, B., Guo, C., Guo, D., Liu, H., Yan, H., Shi, H., Yu, J., Xu, J., Huang, J., et al.: Hy3d-bench: Generation of 3d assets. arXiv preprint arXiv:2602.03907 (2026)
22. Jia, T., Yan, D., Hao, D., Li, Y., Zhang, K., He, X., Li, L., Wang, Y., Chen, J., Jiang, L., et al.: Ultrashape 1.0: High-fidelity 3d shape generation via scalable geometric refinement. arXiv preprint arXiv:2512.21185 (2025)
23. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
24. 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)
25. Lai, Z., Zhao, Y., Zhao, Z., Liu, H., Wang, F., Shi, H., Yang, X., Lin, Q., Huang, J., Liu, Y., et al.: Unleashing vecset diffusion model for fast shape generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2523–2533 (2025)
26. Li, B., Yang, M., Tan, Z., Zhang, J., Li, H.: Unraveling mmdit blocks: Training-free analysis and enhancement of text-conditioned diffusion. arXiv preprint arXiv:2601.02211 (2026)
27. Li, W., Liu, J., Yan, H., Chen, R., Liang, Y., Chen, X., Tan, P., Long, X.: Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. arXiv preprint arXiv:2405.14979 (2024)
28. Li, W., Zhang, X., Sun, Z., Qi, D., Li, H., Cheng, W., Cai, W., Wu, S., Liu, J., Wang, Z., et al.: Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. arXiv preprint arXiv:2505.07747 (2025)
29. 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. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
30. Li, Z., Wang, Y., Zheng, H., Luo, Y., Wen, B.: Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. arXiv preprint arXiv:2505.14521 (2025)
31. Li, Z., Li, H., Shi, Y., Farimani, A.B., Kluger, Y., Yang, L., Wang, P.: Dual diffusion for unified image generation and understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2779–2790 (2025)
32. Liang, Y., Yang, X., Lin, J., Li, H., Xu, X., Chen, Y.: Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. pp. 6517–6526 (2024)
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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 300–309 (2023)
34. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
35. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al.: Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256 (2025)
36. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
37. Pun, A., Deng, K., Liu, R., Ramanan, D., Liu, C., Zhu, J.Y.: Generating physically stable and buildable brick structures from text. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14798–14809 (2025)
38. Ren, X., Huang, J., Zeng, X., Museth, K., Fidler, S., Williams, F.: Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4209–4219 (2024)
39. Shi, Q., Bai, J., Zhao, Z., Chai, W., Yu, K., Wu, J., Song, S., Tong, Y., Li, X., Li, X., et al.: Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. arXiv preprint arXiv:2505.23606 (2025)
40. Siddiqui, Y., Alliegro, A., Artemov, A., Tommasi, T., Sirigatti, D., Rosov, V., Dai, A., Nießner, M.: Meshgpt: Generating triangle meshes with decoder-only transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19615–19625 (2024)
41. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
42. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14398–14409 (2024)
43. Swerdlow, A., Prabhudesai, M., Gandhi, S., Pathak, D., Fragkiadaki, K.: Unified multimodal discrete diffusion. arXiv preprint arXiv:2503.20853 (2025)
44. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
45. 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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22819–22829 (2023)
46. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in neural information processing systems **36**, 1363–1389 (2023)
47. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
48. Team, T.H.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation (2025)

49. Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., Bengio, Y.: Improving and generalizing flow-based generative models with minibatch optimal transport. arXiv preprint arXiv:2302.00482 (2023)
50. Tong, S., Fan, D., Li, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17001–17012 (2025)
51. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
52. Wang, J., Lai, Y., Li, A., Zhang, S., Sun, J., Kang, N., Wu, C., Li, Z., Luo, P.: Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. arXiv preprint arXiv:2505.20147 (2025)
53. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
54. Wang, Z., Lorraine, J., Wang, Y., Su, H., Zhu, J., Fidler, S., Zeng, X.: Llama-mesh: Unifying 3d mesh generation with language models. arXiv preprint arXiv:2411.09595 (2024)
55. 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. Advances in neural information processing systems **36**, 8406–8441 (2023)
56. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
57. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Yang, Y., Bao, Y., Qian, J., Zhu, S., Cao, X., Torr, P., et al.: Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. arXiv preprint arXiv:2505.17412 (2025)
58. Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al.: Vila-u: a unified foundation model integrating visual understanding and generation. arXiv preprint arXiv:2409.04429 (2024)
59. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., et al.: Native and compact structured latents for 3d generation. arXiv preprint arXiv:2512.14692 (2025)
60. 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. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21469–21480 (2025)
61. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
62. 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. pp. 1179–1189 (2023)

63. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809 (2025)
64. 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25050–25061 (2025)
65. Ye, J., Wang, Z., Zhao, R., Xie, S., Zhu, J.: Shapellm-omni: A native multimodal llm for 3d generation and understanding. arXiv preprint arXiv:2506.01853 (2025)
66. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. ACM Transactions On Graphics (TOG) **42**(4), 1–16 (2023)
67. 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. ACM Transactions on Graphics (TOG) **43**(4), 1–20 (2024)
68. Zhang, Y., Zhang, L., Ma, R., Cao, N.: Texverse: A universe of 3d objects with high-resolution textures. arXiv preprint arXiv:2508.10868 (2025)
69. 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)
70. 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. Advances in neural information processing systems **36**, 73969–73982 (2023)
71. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)
72. Zhou, J., Wang, J., Ma, B., Liu, Y.S., Huang, T., Wang, X.: Uni3d: Exploring unified 3d representation at scale. arXiv preprint arXiv:2310.06773 (2023)