Text–Image Conditioned 3D Generation

Jiazhong Cen[1†] Jiemin Fang[2✉] Sikuang Li[1†] Guanjun Wu[3†] Chen Yang[2] Taoran Yi[3†]
Zanwei Zhou[1†] Zhikuan Bao[2] Lingxi Xie[2] Wei Shen[1✉] Qi Tian[2]
[1]MoE Key Lab of Artificial Intelligence, AI Institute,
School of Computer Science, Shanghai Jiao Tong University
[2]Huawei Inc.
[3]Huazhong University of Science and Technology
jaminfong@gmail.com, wei.shen@sjtu.edu.cn

Abstract

High-quality 3D assets are essential for VR/AR, industrial design, and entertainment, motivating growing interest in generative models that create 3D content from user prompts. Most existing 3D generators, however, rely on a single conditioning modality: image-conditioned models achieve high visual fidelity by exploiting pixel-aligned cues but suffer from viewpoint bias when the input view is limited or ambiguous, while text-conditioned models provide broad semantic guidance yet lack low-level visual detail. This limits how users can express intent and raises a natural question: can these two modalities be combined for more flexible and faithful 3D generation? Our diagnostic study shows that even simple late fusion of text- and image-conditioned predictions outperforms single-modality models, revealing strong cross-modal complementarity. We therefore formalize ***Text–Image Conditioned 3D Generation****, which requires joint reasoning over a visual exemplar and a textual specification. To address this task, we introduce* ***TIGON****, a minimalist dual-branch baseline with separate image- and text-conditioned backbones and lightweight cross-modal fusion. Extensive experiments show that text–image conditioning consistently improves over single-modality methods, highlighting complementary vision–language guidance as a promising direction for future 3D generation research. Project page: https://jumpat.github.io/tigon-page*

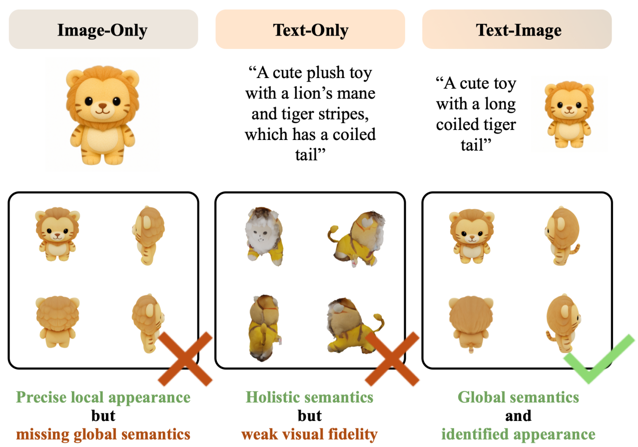

Figure 1. Single-modality conditioning has limitations in satisfying user intent. Image-only conditioning captures local appearance but omits unobserved regions; text-only conveys semantics but lacks visual fidelity. In contrast, joint text–image conditioning produces 3D assets that are both semantically aligned with the description and faithful to the reference appearance.

1. Introduction

Generating high-quality 3D assets has attracted increasing attention due to its importance for downstream applications such as virtual reality, industrial design, and embodied AI. A particularly promising line of work [18, 24, 51, 60, 62, 75] learns to generate 3D objects from either an input image or a text description. However, relying on a single conditioning modality can limit the flexibility of 3D generation.

As shown in Fig. 1, when a user specifies a 3D object with an example image, image-conditioned 3D generation can preserve local appearance but is highly sensitive to viewpoint informativeness: occlusions, atypical views, or incomplete object coverage force the model to hallucinate under-constrained regions, causing the generated 3D asset to deviate from the intended semantics. In contrast, text-conditioned 3D generation is semantically reasonable but lacks concrete visual constraints, so the output may roughly match the prompt while exhibiting poor visual quality.

These observations raise a natural question of whether

†Work done during internship at Huawei.
✉Corresponding authors.

image and text conditions can provide complementary information for more flexible 3D generation. Intuitively, images anchor the result to the actual observed view, supplying reliable geometry and appearance cues, while text can specify additional semantics to disambiguate unobserved regions (*e.g.*, "with a long coiled tiger tail"). In a diagnostic study, we find that conditioning on a low-information view degrades performance, but adding a textual description and fusing the image- and text-conditioned predictions noticeably recovers quality. This motivates us to move beyond single-modality settings and introduce **Text–Image Conditioned 3D Generation**[1], which requires the 3D generator to jointly reason over the visual exemplar and the textual specification, and to generate a consistent 3D asset that is simultaneously faithful to the image-conditioned appearance/geometry and aligned with the text-defined semantics.

To address this task, we propose a strong yet minimalist baseline named **TIGON**. It adopts a dual-branch design that retains two modality-specialized DiT backbones and couples them via two lightweight fusion mechanisms: (i) cross-modal linear bridges for bidirectional feature sharing (early fusion), and (ii) step-wise prediction averaging along the denoising trajectory (late fusion). This design prevents either branch from shouldering the cross-modal domain gap, preserves their original single-modality ability, and enables free-form conditioning. Extensive experiments show that TIGON delivers more flexible 3D generation.

Our contributions are summarized as follows: (1) We identify and empirically diagnose the limitations of existing single-modality 3D generation methods. (2) We show that image and text provide complementary conditioning signals and, motivated by this, introduce the task of text–image conditioned 3D generation. (3) We propose TIGON, a simple yet effective baseline method that leverages modality-specific backbones with lightweight cross-modal fusion. (4) We conduct extensive experiments to demonstrate that TIGON achieves more robust and flexible 3D generation.

2. Related Work

3D Generation with 2D Generative Models. SDS-based text-to-3D methods optimize differentiable 3D representations (*e.g.*, NeRFs [37]) using frozen 2D diffusion priors, as pioneered by DreamFusion [43] and SJC [54]. Later works improve resolution, geometry–appearance disentanglement, optimization stability, and multi-view consistency, while recent hybrids with 3D Gaussian Splatting or native 3D diffusion further boost efficiency and fidelity [4, 20, 23, 25, 27, 28, 47, 50, 55, 61, 72]. A parallel line studies image-conditioned 3D generation and reconstruction with 2D generative priors, either by directly optimizing 3D representations or by first generating consistent multi-view images and then reconstructing 3D content [23, 31–36, 48, 56, 64, 65]. Despite strong progress, reliance on 2D priors often limits 3D consistency, motivating native 3D generative models and cross-modal latent modeling [76].

Among prior works, TICD [14] and FlexGen [67] are most related to ours. TICD augments SDS-based text-to-3D with an image diffusion prior, while FlexGen jointly conditions on text and image but focuses on 2D multi-view generation rather than native 3D synthesis. In contrast, we study **native** 3D generation under joint text-image conditioning.

Native 3D Generative Models. Unlike SDS-based pipelines that rely on 2D generators at test time, native 3D generative models operate directly on 3D representations such as point clouds, meshes, voxels, 3D Gaussians [19, 49], and neural fields [2, 13, 22, 37]. Early works [39, 52] introduce fast point-cloud synthesis via latent diffusion. Later models such as 3DShape2VecSet [73] improve geometry-aware latents for diffusion training. Moreover, scalable voxel-based methods [46], octree-based models [63] and 3DGS-based models [69] improve generation resolution and efficiency.

More recent studies scale data and model capacity for high-quality asset synthesis. Some focus on geometry-only generation for detailed meshes [6, 26, 59, 60, 70], while others target fully textured 3D assets [8, 18, 21, 24, 75, 78]. Among the latter, TRELLIS [62] introduces a sparse, structured latent representation that enables high-fidelity 3D generation via a two-stage pipeline, and UniLat3D [58] provides a unified-latent variant with a simplified sampling process. Together, these advances establish native 3D diffusion models as strong alternatives to 2D diffusion-based pipelines. However, most existing approaches assume single-modality conditioning (either image or text), which limits the flexibility and expressiveness of user instructions. In this work, we study this limitation and introduce the task of text–image conditioned 3D generation.

Multimodal-Conditioned Generation. Multimodal conditioning has proven effective in text, image, and video generation. Multimodal LLMs that ingest both language and visual (or even 3D) inputs [1, 5, 7, 9, 16, 30, 66] enable grounded reasoning and control, while text-to-image diffusion benefits from additional visual controls (edges, depth, pose, layout) and joint text-image inputs [3, 11, 38, 57, 71, 74] for better alignment with user intent. In video generation, combining a reference frame with textual guidance [17, 53] yields more consistent and controllable outputs. These successes motivate our exploration of whether joint text-image conditioning can similarly endow 3D generators with complementary strengths beyond

[1]Here we focus on the *native* 3D generation setting, where the model directly generates a 3D representation under joint text–image conditioning.

single-modality conditioning.

3. Preliminaries

TRELLIS [62] is an effective 3D generator that learns rectified-flow models [12, 29] for a geometry latent (which voxels are active) and an appearance latent defined on those active locations.

Geometry Generation. A 3D object is voxelized into a sparse set of occupied points $\{\mathbf{p}_i\}_{i=1}^N$, $\mathbf{p}_i \in \mathbb{R}^3$. A geometry VAE compresses and reconstructs this set:

$$\mathbf{z}_{\text{geo}} = \mathcal{E}_{\text{geo}}(\{\mathbf{p}_i\}), \qquad \{\mathbf{p}_i\} = \mathcal{D}_{\text{geo}}(\mathbf{z}_{\text{geo}}). \tag{1}$$

A rectified-flow model $\mathcal{F}_{\text{geo}}$ parameterizes a time-conditioned velocity field that transports Gaussian noise toward $\mathbf{z}_{\text{geo}}$ under image or text conditioning.

Appearance Generation. On the occupied positions $\{\mathbf{p}_i\}$, TRELLIS aggregates multi-view image features

$$\mathbf{F} = \{(\mathbf{p}_i, \mathbf{f}_i)\}_{i=1}^N, \quad \mathbf{f}_i \in \mathbb{R}^d,$$

extracted by a 2D vision encoder such as DINOv2 [40], and encodes them into a Structured LATent (SLAT) with an appearance VAE:

$$\mathbf{z}_{\text{SLAT}} = \mathcal{E}_{\text{app}}(\mathbf{F}) = \{(\mathbf{p}_i, \mathbf{z}_i)\}_{i=1}^N. \tag{2}$$

A second rectified-flow model $\mathcal{F}_{\text{app}}$ is learned for $\mathbf{z}_{\text{SLAT}}$, and the appearance decoder produces the final 3D representation

$$\mathcal{O} = \mathcal{D}_{\text{app}}(\mathbf{z}_{\text{SLAT}}), \tag{3}$$

where $\mathcal{O}$ can be mesh, 3DGS [19], or radiance fields [13].

Sampling with Rectified Flow. For each latent $\ell \in \{\text{geo}, \text{SLAT}\}$ with decoder $\mathcal{D}_\ell \in \{\mathcal{D}_{\text{geo}}, \mathcal{D}_{\text{app}}\}$ and a chosen condition $\mathbf{c}$ (image $\mathbf{I}$ or text $\mathbf{T}$), inference integrates the rectified-flow ODE from noise to data. Let $1 = t_0 > t_1 > \cdots > t_K = 0$ be a fixed schedule and initialize $\tilde{\mathbf{z}}_{\ell,t_0} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}^c)$, where c is the latent dimension. At step k,

$$\begin{aligned} \mathbf{v}_{\ell,k} &= \mathcal{F}_\ell(\tilde{\mathbf{z}}_{\ell,t_k}, t_k; \mathbf{c}), \\ \tilde{\mathbf{z}}_{\ell,t_{k+1}} &= \tilde{\mathbf{z}}_{\ell,t_k} - (t_k - t_{k+1})\,\mathbf{v}_{\ell,k}, \end{aligned} \tag{4}$$

yielding $\tilde{\mathbf{z}}_{\ell,0}$ after K steps. Decoding $\tilde{\mathbf{z}}_{\text{geo},0}$ with $\mathcal{D}_{\text{geo}}$ produces the activated voxels, and decoding $\tilde{\mathbf{z}}_{\text{SLAT},0}$ with $\mathcal{D}_{\text{app}}$ gives the final 3D output $\mathcal{O}$.

Figure 2. Reference views used in our diagnostic study. Moving from *View-0* to *View-1* reduces observable cues and creates a lower-information setting. Under this shift, single-modality baselines exhibit a marked performance drop.

UniLat3D. Based on TRELLIS, UniLat3D [58] offers a more convenient single-stage 3D generator. Given view-aggregated features $\mathbf{F}$, it encodes them as

$$\mathbf{z}_{\text{uni}} = \mathcal{E}_{\text{uni}}(\mathbf{F}), \quad \mathbf{z}_{\text{uni}} \in \mathbb{R}^{16\times16\times16\times c}, \tag{5}$$

and directly decodes to a 3D output $\mathcal{O} = \mathcal{D}_{\text{uni}}(\mathbf{z}_{\text{uni}})$. A single rectified-flow model predicts $\tilde{\mathbf{z}}_{\text{uni}}$ during generation. We adopt UniLat3D for its simplicity and compatibility with our conditioning study.

4. Text-Image Conditioned 3D Generation

As discussed in Sec. 1, image-conditioned 3D generation is vulnerable to viewpoint bias, while text-conditioned generation benefits from comprehensive semantics but lacks the visual cues needed for high-fidelity synthesis. In this section, we first empirically diagnose this limitation, then show that even a simple late fusion of image- and text-conditioned predictions yields noticeable gains, revealing clear complementarity between the two modalities and motivating the task of text–image conditioned 3D generation.

Limitations of Single-Modality Conditioning. Tab. 1 reports the performance of representative 3D generation models on Toys4K under two viewpoint configurations[2]. As illustrated in Fig. 2, *View-0* is a frontal view with rich semantics and clear local details, whereas *View-1* is a low-angle view providing much weaker cues. This change alone leads to substantial degradation: TRELLIS degrades from 56.08 $\text{FD}_{\text{DINOv2}}$ under *View-0* to 143.58 under *View-1*, indicating strong dependence on viewpoint completeness. Moreover, the text-only counterparts perform even worse in visual alignment (*e.g.*, UniLat3D reaches only 154.88 $\text{FD}_{\text{DINOv2}}$), confirming that text priors alone are insufficient to recover fine-grained visual details.

Enhanced 3D Generation with Simple Cross-Modal Fusion. Images and text provide complementary constraints:

[2]Please refer to Sec. 6.2 for details about the metrics used in Tab. 1.

Table 1. Performance of existing methods on the Toys4K dataset under different conditioning signals. 'GS' denotes that the 3D representation is 3DGS.

Model	Cond.	CLIP↑	FD_{DINOv2} ↓
TripoSR	View-0	88.67	269.58
Step1X-3D†	View-0	89.99	152.69
Hunyuan3D-2.1†	View-0	89.87	114.64
TRELLIS (GS)	View-0	92.88	56.08
UniLat3D (GS)	View-0	93.34	47.41
TripoSR	View-1	79.40	804.18
Step1X-3D†	View-1	80.47	562.84
Hunyuan3D-2.1†	View-1	85.33	229.36
TRELLIS (GS)	View-1	88.16	143.58
UniLat3D (GS)	View-1	89.03	125.93
TRELLIS (GS)	Text	86.30	148.21
UniLat3D (GS)	Text	86.14	154.88
SimFusion (GS; Ours)	View-1 + Text	**90.64**	**82.40**

† Using non-public training data.

text offers high-level, multi-view semantic priors, while images provide precise cues on style, texture, geometry, and color. To verify this complementarity, we conduct a simple fusion experiment. At inference time, we take two pretrained rectified-flow models, one image-conditioned and the other text-conditioned, and directly average their predicted velocity fields at each denoising step to form a joint text-image baseline, which we call SimFusion.

As shown in Tab. 1, this naive fusion already outperforms both image-only and text-only models by a large margin (82.40 FD vs. 125.93 and 145.06), suggesting that it preserves the semantic correctness from text while retaining the fine-grained visual cues from the image. This complementary effect motivates us to define a new task, termed **Text–Image Conditioned 3D Generation**.

Problem Formulation. In text-image conditioned 3D generation, the model should jointly adopt a visual exemplar (image) and a semantic description (text) to generate a coherent 3D object. Formally, given an image condition $\mathbf{I}$ and a text condition $\mathbf{T}$, the goal is to model the conditional distribution $p(\mathcal{O} \mid \mathbf{I}, \mathbf{T})$, where $\mathcal{O}$ denotes the target 3D representation (*e.g.*, mesh, 3DGS, or radiance field). This task requires the model to (i) satisfy the semantics in $\mathbf{T}$ and (ii) match the view-specific appearance constraints in $\mathbf{I}$.

5. Method

In this section, we introduce **TIGON** (Text–Image conditioned GeneratiON) as a baseline for our proposed task.

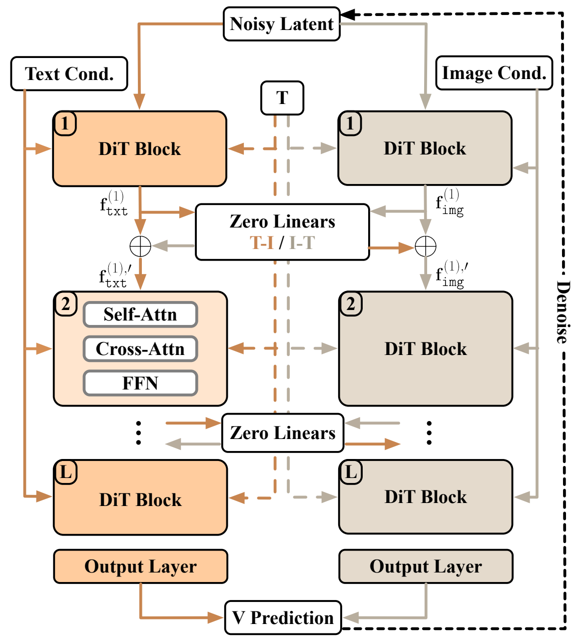

Figure 3. TIGON employs a dual-branch architecture, with a text-conditioned DiT (left) and an image-conditioned DiT (right). Paired blocks exchange features via cross-modal bridges ("Zero Linears"). At each denoising step, two predictions are averaged to produce the velocity field $\mathbf{v}$. T denotes the denoising timestep.

5.1. Overall Pipeline

As shown in Fig. 3, **TIGON** uses two parallel branches (image- and text-conditioned) that exchange features via zero-initialized cross-modal bridges between corresponding DiT blocks. At each denoising step, their predictions are averaged to produce the final velocity field $\mathbf{v}$.

5.2. Dual-Branch Backbone

TIGON adopts a dual-branch backbone because image and text provide fundamentally different signals for 3D generation. Image-conditioned tokens are dense, view-grounded, and locally informative, offering explicit cues about color, texture, and fine geometry, whereas text-conditioned tokens encode sparse, abstract semantics. This asymmetry creates a granularity mismatch: for example, the concept "tiger" may be conveyed by a single word token but requires many image tokens to depict. Without enough data, mixing such heterogeneous token semantics within a single backbone often degrades performance. Therefore, TIGON retains two modality-specific backbones and performs fusion explicitly, preserving each branch's strengths while avoiding overly aggressive entanglement.

Each branch of TIGON is a Diffusion Transformer (DiT) [42] with L blocks. Let $\mathcal{F}_{\mathtt{img}}$ and $\mathcal{F}_{\mathtt{txt}}$ denote the image- and text-conditioned branches. Given latent $\tilde{\mathbf{z}}$, time step t, and condition (image $\mathbf{I}$ or text $\mathbf{T}$), each branch predicts a velocity field:

$$\mathbf{v}_{\mathtt{img}} = \mathcal{F}_{\mathtt{img}}(\tilde{\mathbf{z}},\, t,\, \mathbf{I})\,,$$
$$\mathbf{v}_{\mathtt{txt}} = \mathcal{F}_{\mathtt{txt}}(\tilde{\mathbf{z}},\, t,\, \mathbf{T})\,. \tag{6}$$

Both branches are pretrained in the same latent space introduced in Sec. 3, which allows simple additive fusion of the predicted velocities.

5.3. Early-Fusion Strategy

Simply averaging the final predictions of two rectified flow models is often sub-optimal: without explicit interaction, the branches can diverge and destructive averaging degrades detail and consistency. We therefore assign a **cross-modal bridge** at every backbone block for early, fine-grained cross-modal feature fusion.

Let the image- and text-conditioned branches each have L blocks. Denote by $\mathbf{f}^{(i)}_{\mathtt{img}}$ and $\mathbf{f}^{(i)}_{\mathtt{txt}}$ the output of the i-th block (i=$1,\ldots,L$). We insert learned linear projections $\mathcal{P}^{(i)}_{\mathtt{txt}\to\mathtt{img}}$ and $\mathcal{P}^{(i)}_{\mathtt{img}\to\mathtt{txt}}$ to inject information across branches. The inputs to the (i+1)-th blocks are:

$$\mathbf{f}^{(i),\prime}_{\mathtt{img}} = \mathbf{f}^{(i)}_{\mathtt{img}} \;+\; \mathcal{P}^{(i)}_{\mathtt{txt}\to\mathtt{img}}\big(\mathbf{f}^{(i)}_{\mathtt{txt}}\big),$$
$$\mathbf{f}^{(i),\prime}_{\mathtt{txt}} = \mathbf{f}^{(i)}_{\mathtt{txt}} \;+\; \mathcal{P}^{(i)}_{\mathtt{img}\to\mathtt{txt}}\big(\mathbf{f}^{(i)}_{\mathtt{img}}\big). \tag{7}$$

Stability via Zero-Initialization Inspired by ControlNet [74], to maintain training stability at the start of joint training, all cross-modal bridges are zero-initialized. Consequently, $\mathbf{f}^{(i),\prime}_{\mathtt{img}} = \mathbf{f}^{(i)}_{\mathtt{img}}$ and $\mathbf{f}^{(i),\prime}_{\mathtt{txt}} = \mathbf{f}^{(i)}_{\mathtt{txt}}$ initially, and gradients progressively "open" these gates, learning when (and how much) to exchange information at each depth.

5.4. Late-Fusion Strategy

We adopt a simple prediction-averaging scheme. Given the outputs of the text and image branches, $\mathbf{v}_{\mathtt{txt}}$ and $\mathbf{v}_{\mathtt{img}}$, the fused prediction at each denoising step is

$$\mathbf{v} = \tfrac{1}{2}\big(\mathbf{v}_{\mathtt{txt}} + \mathbf{v}_{\mathtt{img}}\big). \tag{8}$$

Why Is a Sophisticated Fusion Strategy Unnecessary? Early fusion with cross-modal bridges and end-to-end fine-tuning enable each branch to implicitly condition on both modalities, so any potential benefit of dynamic, modality-weighted fusion can be absorbed into the branch parameters (*i.e.*, by reparameterization during training). To validate this, we compare against two learnable late-fusion variants: (i) a weight-prediction module that outputs a scalar for linear mixing; and (ii) an additional cross-modal attention block. Both yield at most marginal gains while introducing extra parameters and training variance. See Sec. 6.5 for quantitative results; architectural details are in the supplementary material.

5.5. Training Strategy

The training of TIGON is a two-stage process. The two branches are first pre-trained separately on their respective modalities to ensure balanced learning. Then, the zero-initialized cross-modal bridges are trained, and all model parameters are jointly fine-tuned.

To preserve unimodal generation capability, we apply condition dropout: during training, the image and text conditions are independently dropped with probability 0.5. This produces a uniform mixture over four regimes—25% unconditional (for CFG [15]), 25% text-only, 25% image-only, and 25% text+image. Consequently, TIGON learns to handle **free-form conditioning** at inference, supporting text-only, image-only, or joint text–image inputs.

6. Experiments

We first describe the implementation details, datasets, and evaluation protocol, then report quantitative and qualitative results and ablations.

6.1. Implementation Details

TIGON is implemented in PyTorch [41] on top of TRELLIS and UniLat3D. We use the released UniLat3D checkpoint for the image branch. For the text branch, we reuse the UniLat3D backbone, replace its DINO-based condition encoder with a CLIP text encoder, and train from scratch for $1{,}000{,}000$ iterations with batch size 256 and learning rate 1×10^{-4}. We then jointly fine-tune both branches and the cross-modal bridges for $50{,}000$ iterations with learning rate 1×10^{-5} in BF16 on 64 NVIDIA A800 GPUs, using DeepSpeed ZeRO-2 [45] and FlashAttention [10].

6.2. Datasets and Evaluation Protocol

Training. TIGON is trained on **TRELLIS-500K**. Please refer to the supplement for more details about this dataset.
Evaluation. We evaluate on two test sets: extbfToys4K contains about 4K high-quality 3D objects from 105 categories and is widely used by TRELLIS and UniLat3D. **UniLat1K** is a harder 1K-object benchmark curated by UniLat3D, containing 500 high-quality Sketchfab assets and 500 Toys4K samples.
Metrics and Protocol. We use four metrics: CLIP [44], $\mathrm{FD}_{\mathrm{DINOv2}}$, ULIP [68], and Uni3D [77]. CLIP and $\mathrm{FD}_{\mathrm{DINOv2}}$ are computed from renderings of generated and ground-truth objects, while ULIP and Uni3D measure image–point-cloud alignment and are thus only reported for mesh outputs. To test robustness to viewpoint informativeness, each case is conditioned on three reference views (front, top, and bottom) instead of ideal views. We use public checkpoints for prior methods and re-evaluate them under this unified protocol; full metric definitions and rendering settings are provided in the supplement.

Table 2. Quantitative results on Toys4K (left) and UniLat1K (right). “Cond.” denotes conditioning modality (“I”: image, “T”: text), and “Rep.” denotes output representation (“M.”: mesh, “GS”: 3DGS).

Model	Cond.	Rep.	Toys4K				UniLat1K			
			CLIP↑	FD_{DINOv2} ↓	ULIP↑	Uni3D↑	CLIP↑	FD_{DINOv2} ↓	ULIP↑	Uni3D↑
TripoSR [51]	I	M.	83.14	596.44	27.37	24.38	83.37	652.27	25.90	23.96
TRELLIS	I	M.	89.09	171.44	39.97	35.61	89.40	233.53	39.37	35.40
TRELLIS	I	GS	90.50	98.75	-	-	90.83	177.20	-	-
Step1X-3D† [24]	I	M.	84.77	361.44	34.15	30.04	85.36	402.25	33.62	30.38
Hunyuan3D-2.1† [18]	I	M.	87.57	171.91	40.22	35.70	87.27	249.66	39.58	35.54
Stable3DGen [70]	I	M.	-	-	35.52	31.76	-	-	35.26	32.08
Direct3D-S2 [60]	I	M.	-	-	33.47	29.29	-	-	32.74	29.44
UniLat3D	I	M.	91.85	109.68	40.32	35.75	90.00	205.72	39.60	35.49
UniLat3D	I	GS	91.20	85.30	-	-	91.40	155.99	-	-
TIGON (Ours)	I	GS	91.40	84.62	-	-	91.64	153.79	-	-
TRELLIS	T	M.	87.15	182.42	37.41	33.59	85.90	316.05	36.55	33.23
TRELLIS	T	GS	86.30	148.21	-	-	84.75	288.55	-	-
UniLat3D	T	M.	87.03	179.95	36.14	32.35	85.29	313.85	35.34	32.08
UniLat3D	T	GS	86.14	154.88	-	-	85.75	282.36	-	-
TIGON (Ours)	T	GS	86.77	152.34	-	-	86.42	273.97	-	-
SimFusion (Ours)	I+T	GS	91.95	66.78	-	-	92.09	136.97	-	-
TIGON (Ours)	I+T	M.	**92.97**	80.77	**41.36**	**36.68**	90.91	176.69	**40.95**	**36.74**
TIGON (Ours)	I+T	GS	92.33	**61.59**	-	-	**92.42**	**130.08**	-	-

† Using non-public training data.

6.3. Quantitative Results

Results on Toys4K and UniLat1K are reported in Table 2. We evaluate TIGON under three conditioning regimes, *i.e.*, text-only, image-only, and text–image. TIGON is competitive in both single-modality settings, while the largest gains appear under text–image conditioning, showing effective use of complementary signals.

6.4. Qualitative Results

As shown in Fig. 4, we compare TIGON with image-only and text-only variants of TRELLIS and UniLat3D.

Image-Only Conditioning. With only one reference view, shape and appearance remain under-constrained, so image-only models must hallucinate unseen regions and often deviate from user intent. For example, given a top view of a trophy, TRELLIS/UniLat3D capture the overall style but fail to reconstruct a faithful trophy. Even with a more informative view (*e.g.*, the toaster in the last row), they still produce distorted slots due to incomplete observation. Adding text supplies the missing semantics, so TIGON better matches both the description and the reference image.

Text-Only Conditioning. Text provides high-level semantics but no pixel-aligned cues, leading to ambiguous geometry and lower visual fidelity. Introducing even a weak image cue markedly improves spatial alignment and appearance: in the third row of Fig. 4, a top-view image of a game console combined with text produces much more faithful geometry than the text-only baseline. Overall, the visual results show that text and image address complementary failure modes, and that joint conditioning enables more controllable, higher-quality 3D generation.

To further demonstrate TIGON's controllability, we fix the input image and vary the text prompt to obtain diverse 3D outputs. As shown in Fig. 5, when the image is highly informative (*e.g.*, a distinctive character), TIGON preserves identity while adjusting fine-grained attributes according to the text; when the image is ambiguous (*e.g.*, a bottom view of a toaster), it relies more on text and can generate semantically different objects, such as a telephone or a bed. This combination of pixel-level alignment and semantic control offers greater flexibility than single-modality methods. We further observe in Fig. 6 that when the image and text explicitly conflict, TIGON tends to follow the image if it already provides clear semantic guidance, likely because images are usually more specific and less ambiguous than text. More qualitative results are provided in the supplement.

6.5. Ablation Study

We study each TIGON component on Toys4K.

Early-Fusion Strategy. As shown in Tab. 3, without cross-modal bridges, joint fine-tuning of the two branches

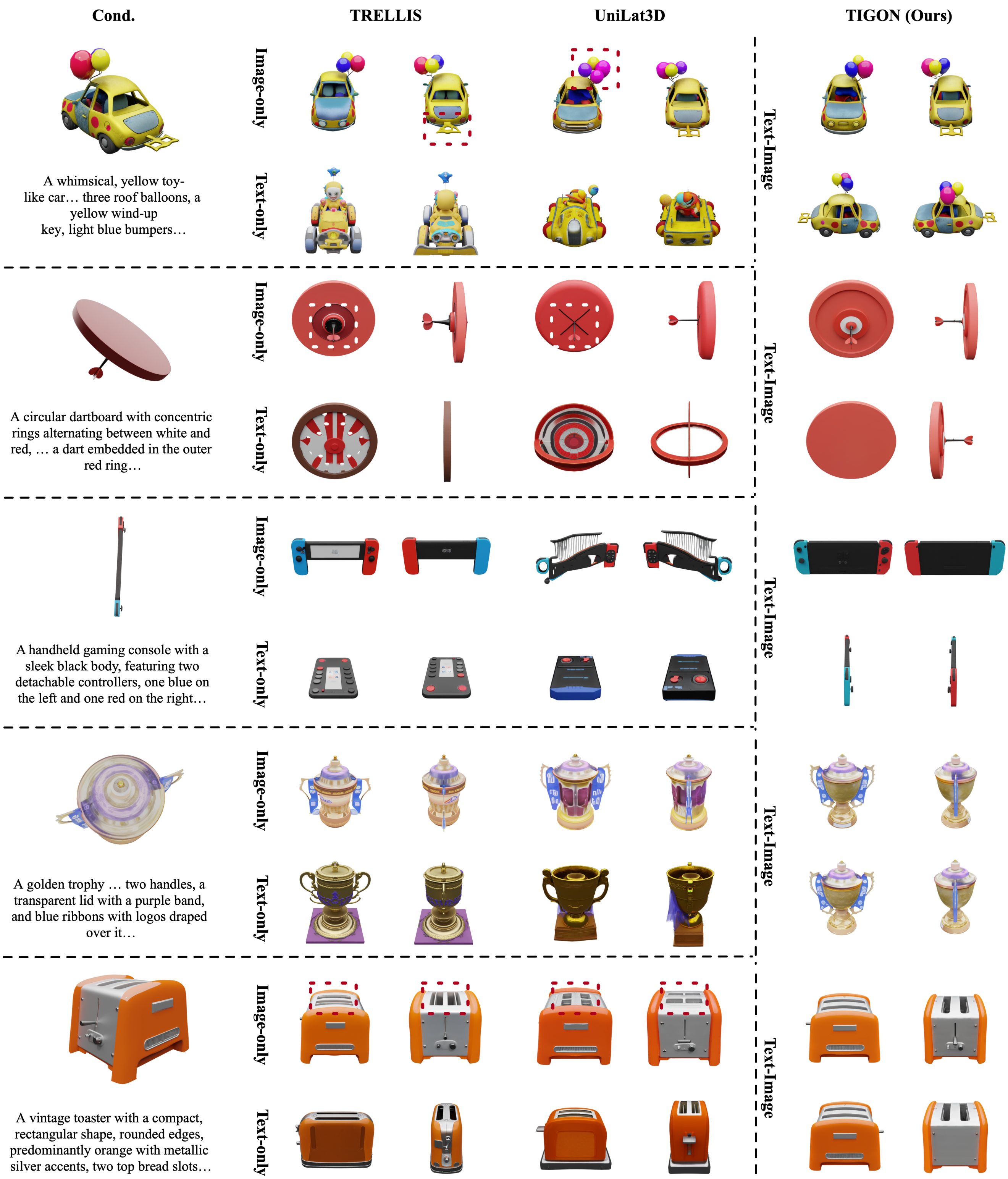

Figure 4. Qualitative comparison on Toys4K and UniLat1K against image-only and text-only variants of TRELLIS and UniLat3D. Dashed boxes mark artifacts from prior methods. Image-only models respect the reference view but must hallucinate unseen regions, while text-only models lack pixel-aligned cues and often produce low-fidelity geometry and appearance. Full prompts are provided in the supplement.

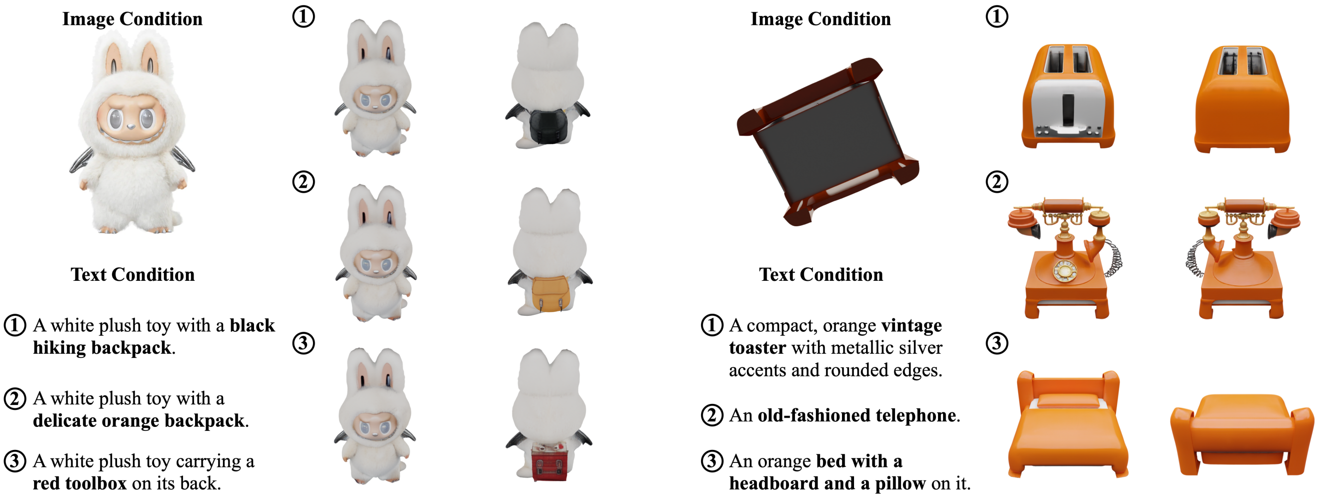

Figure 5. Controllable generation under text and image conditions.

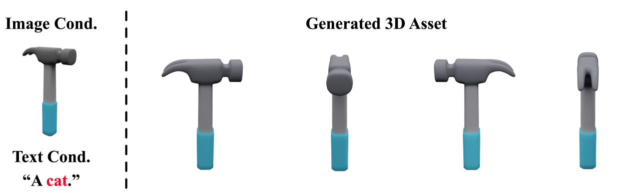

Figure 6. Generation with conflicting text-image conditions.

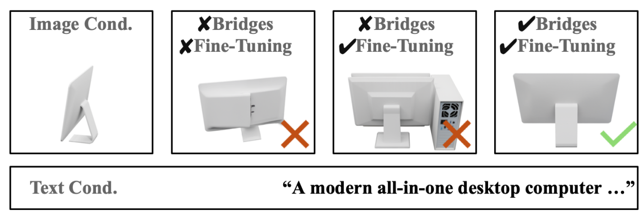

Figure 7. Effect of early fusion. Without cross-modal bridges, the two branches diverge during denoising. Full text prompt is available in the supplement.

Table 3. Ablations on Toys4K. "Bridges" denotes zero-initialized cross-modal bridges; "Sim", "AW", and "AT" denote three fusion strategies; "FT" denotes joint fine-tuning. The TIGON setting is highlighted in light gray.

Bridges	Fusion Strategy			FT	CLIP↑	FD_{DINOv2} ↓
	Sim	AW	AT			
	✓				91.95	66.78
	✓			✓	92.05	66.04
✓	✓			✓	92.33	61.59
✓		✓		✓	92.31	60.90
✓			✓	✓	92.26	62.00

only brings marginal improvement ($66.78 \rightarrow 66.04$ in FD_{DINOv2}). Enabling cross-modal bridges yields a substantial gain ($66.78 \rightarrow 61.59$ in FD_{DINOv2}), underscoring the necessity of cross-modal information exchange. Qualitatively, Fig. 7 shows that, without cross-modal bridges, the text- and image-conditioned branches diverge during denoising, producing inconsistent or abnormal structures; with bridges, they remain aligned and produce coherent result.

Late-Fusion Strategy. As discussed in Sec. 5.4, a sophisticated learnable late-fusion strategy appears unnecessary. Under the same early-fusion setup and joint fine-tuning, simple averaging (**Sim**) already reaches 61.59 FD_{DINOv2} (Table 3), while adaptive weighting (**AW**) and attention-based fusion (**AT**) change this metric only slightly (60.90 and 62.00). We therefore adopt simple averaging by default.

7. Conclusion

In this paper, we revisit conventional single-modality conditioned 3D generation and highlight a clear limitation: image-conditioned models are sensitive to viewpoint informativeness and lack control over unobserved regions, while text-conditioned models capture global intent but lack concrete visual cues. Our diagnostic study empirically shows that these two signals are complementary, which motivates us to formalize the task of **Text–Image Conditioned 3D Generation**. We further introduce **TIGON**, a simple yet effective baseline for this task. Experiments demonstrate consistent gains over strong single-modality models, and show that combining modalities yields more robust and flexible 3D generation. We hope this work will help drive future research on controllable, high-quality 3D generation.

Acknowledgements

This work was supported by the NSFC under Grant 62322604 and 62576207. We thank the anonymous reviewers for their valuable feedback and suggestions.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 2

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2

[4] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023. 2

[5] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *CVPR*, 2024. 2

[6] Yiwen Chen, Zhihao Li, Yikai Wang, Hu Zhang, Qin Li, Chi Zhang, and Guosheng Lin. Ultra3d: Efficient and high-fidelity 3d generation with part attention. *arXiv preprint arXiv:2507.17745*, 2025. 2

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 2

[8] Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In *CVPR*, 2025. 2

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 2023. 2

[10] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 5

[11] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2

[12] 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 *ICLR*, 2024. 3

[13] Quankai Gao, Qiangeng Xu, Hao Su, Ulrich Neumann, and Zexiang Xu. Strivec: Sparse tri-vector radiance fields. In *ICCV*, 2023. 2, 3

[14] Yuze He, Yushi Bai, Matthieu Lin, Jenny Sheng, Yubin Hu, Qi Wang, Yu-Hui Wen, and Yong-Jin Liu. Text-image conditioned diffusion for consistent text-to-3d generation. *Computer Aided Geometric Design*, 2024. 2

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[16] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *NeurIPS*, 2023. 2

[17] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *CVPR*, 2022. 2

[18] 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. 1, 2, 6

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 2, 3

[20] Min-Seop Kwak, Donghoon Ahn, Inès Hyeonsu Kim, Jin-Hwa Kim, and Seungryong Kim. Geometry-aware score distillation via 3d consistent noising and gradient consistency modeling. *arXiv preprint arXiv:2406.16695*, 2024. 2

[21] 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. 2

[22] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *ECCV*, 2024. 2

[23] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 2

[24] 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. 1, 2, 6

[25] Zongrui Li, Minghui Hu, Qian Zheng, and Xudong Jiang. Connecting consistency distillation to score distillation for text-to-3d generation. In *ECCV*, 2024. 2

[26] Zhihao Li, Yufei Wang, Heliang Zheng, Yihao Luo, and Bihan Wen. Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. *arXiv preprint arXiv:2505.14521*, 2025. 2

[27] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *CVPR*, 2024. 2

[28] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 2

[29] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[31] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *NeurIPS*, 2023. 2

[32] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *CVPR*, 2024.

[33] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023.

[34] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *ICLR*, 2024.

[35] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, 2024.

[36] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *CVPR*, 2023. 2

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[38] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 2

[39] 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. 2

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 3

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 5

[42] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4

[43] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, 2021. 5

[45] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020. 5

[46] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *CVPR*, 2024. 2

[47] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. In *ICLR*, 2024. 2

[48] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *ICCV*, 2023. 2

[49] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *ECCV*, 2024. 2

[50] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024. 2

[51] 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. 1, 6

[52] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *NeurIPS*, 2022. 2

[53] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi

Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2

[54] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023. 2

[55] 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. *NeurIPS*, 2023. 2

[56] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. 2

[57] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2

[58] Guanjun Wu, Jiemin Fang, Chen Yang, Sikuang Li, Taoran Yi, Jia Lu, Zanwei Zhou, Jiazhong Cen, Lingxi Xie, Xiaopeng Zhang, et al. Unilat3d: Geometry-appearance unified latents for single-stage 3d generation. *arXiv preprint arXiv:2509.25079*, 2025. 2, 3

[59] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems*, 2024. 2

[60] 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. *arXiv preprint arXiv:2505.17412*, 2025. 1, 2, 6

[61] Zike Wu, Pan Zhou, Xuanyu Yi, Xiaoding Yuan, and Hanwang Zhang. Consistent3d: Towards consistent high-fidelity text-to-3d generation with deterministic sampling prior. In *CVPR*, 2024. 2

[62] 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 *CVPR*, 2025. 1, 2, 3

[63] Bojun Xiong, Si-Tong Wei, Xin-Yang Zheng, Yan-Pei Cao, Zhouhui Lian, and Peng-Shuai Wang. Octfusion: Octree-based diffusion models for 3d shape generation. *Computer Graphics Forum*, 2025. 2

[64] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *CVPR*, 2023. 2

[65] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2

[66] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024. 2

[67] Xinli Xu, Wenhang Ge, Jiantao Lin, Jiawei Feng, Lie Xu, HanFeng Zhao, Shunsi Zhang, and Ying-Cong Chen. Flexgen: Flexible multi-view generation from text and image inputs. In *ICCV*, 2025. 2

[68] 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 *CVPR*, 2023. 5

[69] Haitao Yang, Yuan Dong, Hanwen Jiang, Dejia Xu, Georgios Pavlakos, and Qixing Huang. Atlas gaussians diffusion for 3d generation. In *ICLR*, 2025. 2

[70] 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. *arXiv preprint arXiv:2503.22236*, 2025. 2, 6

[71] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 2

[72] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, 2024. 2

[73] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM TOG*, 2023. 2

[74] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 5

[75] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM TOG*, 2024. 1, 2

[76] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *NeurIPS*, 2023. 2

[77] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 5

[78] Zanwei Zhou, Taoran Yi, Jiemin Fang, Chen Yang, Lingxi Xie, Xinggang Wang, Wei Shen, and Qi Tian. Few-step flow for 3d generation via marginal-data transport distillation. In *AAAI*, 2026. 2