OMNI123: EXPLORING 3D NATIVE FOUNDATION MODELS WITH LIMITED 3D DATA BY UNIFYING TEXT TO 2D AND 3D GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models excel at text and image generation, but native 3D generation and editing remain constrained by scarce high-quality 3D data. Existing approaches often operate in 2D and lift results into 3D, compromising geometric consistency. We present Omni123, a native 3D foundation model that unifies text-to-2D and text-to-3D generation in a single autoregressive framework. Our key insight is that cross-modal generative consistency can provide an implicit structural constraint, enabling abundant 2D observations to strengthen 3D representations. Omni123 encodes text, images, and 3D geometry as discrete tokens and uses interleaved X-to-X training over heterogeneous paired datasets, without requiring aligned text–image–3D triplets. Semantic–visual–geometric cycles, such as text $\rightarrow$ image $\rightarrow$ 3D $\rightarrow$ image, coordinate semantic intent, appearance fidelity, and multi-view consistency while mitigating interference between appearance and geometry objectives. Experiments demonstrate improved geometric consistency and semantic alignment in text-guided 3D generation and editing, supporting unified 2D–3D generation as a scalable route toward multimodal 3D world models.

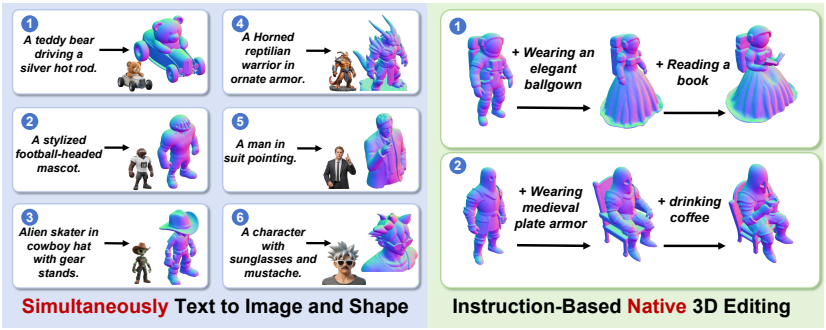

Figure 1: Omni123 for native 3D generation and editing. Left: Jointly generated images and 3D shapes (normal maps). Right: Sequential text-guided 3D edits.

1 INTRODUCTION

Unified multimodal models, including GPT-4o (OpenAI, 2025) and Nano-Banana (Google, 2026), enable native high-fidelity image synthesis and editing within a shared framework. Extending these capabilities to a *3D Nano-Banana* could advance embodied AI and autonomous virtual world synthesis, yet native 3D generation and editing remain challenging. Unlike 2D AIGC with abundant web-scale data, 3D generation faces acute data scarcity. This disparity leaves 3D synthesis inherently under-constrained, especially under sparse supervision from text. As shown in Fig. 3(a), scaling text $\rightarrow$ 3D training yields continued gains but soon hits a bottleneck due to limited accessible data, suggesting that 3D learning requires stronger structural constraints and broader semantic knowledge. Crucially, 2D observations implicitly encode 3D structural information, including object shape (Lin et al., 2025), appearance (Wang et al., 2025), and spatial relationships (Baik et al., 2025). This motivates a pivotal question: Can we leverage the abundance of 2D observations to fortify 3D generation in a unified multimodal framework? However, joint training alone does not guarantee

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

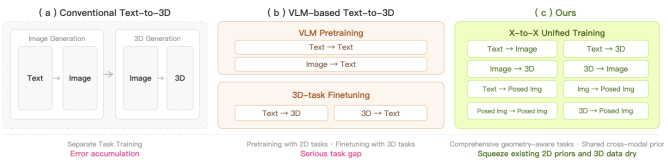

Figure 2: Comparison of text-to-3D paradigms.

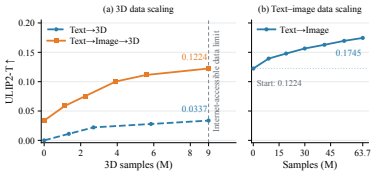

Figure 3: Scaling with 3D and text–image data for text-to-3D.

effective knowledge transfer. These tasks rely on different priors: text-to-image generation models appearance distributions, whereas text-to-3D generation requires geometric consistency. Without careful coordination, competing gradients may cause harmful task interference and degrade unified representations.

To address this challenge, we introduce Omni123, a unified autoregressive model built on an **interleaved X-to-X training paradigm**. **We posit that cross-modal generative consistency between images and 3D can act as an implicit structural constraint for learning robust 3D representations.** These modalities provide complementary observations of the same object: text provides semantic intent, images capture appearance and partial geometry, and 3D provides explicit spatial structure. We encode these complementary signals as discrete tokens in a shared sequence space and train the model on heterogeneous paired datasets, allowing any modality to serve as input or target without requiring fully complete triplets. Interleaving these modalities in a semantic–visual–geometric cycle (e.g., text $\rightarrow$ image $\rightarrow$ 3D $\rightarrow$ image) encourages representations to simultaneously preserve semantic intent, appearance fidelity, and multi-view consistency. Fig. 3(a) shows that, with the same number of 3D assets, interleaved training substantially improves performance and exhibits better scaling behavior than pure text-to-3D training. This motivates us to further incorporate general-domain text-to-image data, bringing broader concepts, appearances, and spatial relationships to 3D learning. Fig. 3(b) shows a power-law-like improvement as general text-to-image data scales, suggesting that **2D observations offer a more efficient and promising path for scaling 3D generation**.

Beyond generation, a *3D Nano-Banana* should support faithful instruction-based editing while preserving non-edited geometry. Using a novel scalable data engine, we synthesized **3DEdit-1M**, a 3D editing dataset with 1M aligned image–shape pairs and imaginative, open-ended instructions. We also introduce **Consis3D-Bench**, an artist-modeled benchmark with precise edits and strictly preserved non-edited regions for reliable consistency evaluation. We fine-tune Omni123 on 3DEdit-1M, achieving SOTA editing performance and enabling sequential edits in Fig. 1.

Experiments demonstrate that interleaved cross-modal generation significantly improves geometric consistency and semantic alignment in text-guided 3D generation and editing tasks. These findings suggest that cross-modal generative consistency provides an effective pathway toward scalable multimodal 3D world models. In summary, our key contributions include:

- We introduce Omni123, a shared-token autoregressive model that unifies text, images, and 3D, supporting native 3D generation and sequential instruction-based editing without intermediate 2D editing proxies or iterative image-space optimization.
- We propose an interleaved training scheme that coordinates diverse generative tasks across heterogeneous datasets. This approach mitigates task interference between appearance and geometry, facilitating beneficial cross-modal knowledge transfer from 2D to 3D.
- We leverage generative consistency across "semantic-visual-geometric" cycles as an implicit structural constraint. This allows abundant 2D observations to serve as a rich geometric prior, effectively fortifying 3D representations against acute data scarcity.
- We build a 1M-example 3D editing dataset with imaginative, open-ended instructions, paired with an artist-modeled benchmark for faithful 3D editing evaluation.

2 Related Work

2.1 3D Generation

3D generation has advanced rapidly, from distilling 2D diffusion priors into 3D (Poole et al., 2022; Qian et al., 2023; Qiu et al., 2024) to 3D-native pipelines that learn 3D distributions from large-scale

datasets (Xiang et al., 2025b;a; Hunyuan3D Team et al., 2025). Previous works design various 3D representations with different trade-offs between geometric detail, computational cost, and scalability, including point clouds (Nichol et al., 2022), voxels (Xiong et al., 2025a), meshes (Liu et al., 2023), Gaussian primitives (Lan et al., 2025), and tri-planes (Wu et al., 2024). For compact generative modeling, 3DShape2VecSet (Zhang et al., 2023) encodes shapes as sets of latent vectors, subsequently scaled by CLAY (Zhang et al., 2024) and TripoSG (Li et al., 2025b). TRELLIS (Xiang et al., 2025b) steps forward to jointly encode geometry and appearance in sparse voxels, and Direct3D-S2 (Wu et al., 2025) reduces the cost of high-resolution generation through spatial sparse attention. Despite these advances, current 3D generation primarily focuses on image-to-3D generation. Direct text-to-3D remains challenging due to limited geometric guidance in text and a larger modality gap. Our cross-modality interleaved training paradigm bridges language and geometry through images, transferring visual priors to text-conditioned 3D generation.

2.2 3D Editing

Generative models increasingly support 3D editing (Sella et al., 2023). VoxHammer (Li et al., 2025a) uses inversion and 3D masks, while EditFlow3D (Nie et al., 2026) uses inversion-free differential flow. Prox-E (Sella et al., 2026) enables control through primitive-based proxies, and Nano3D (Ye et al., 2026b) extends FlowEdit (Kulikov et al., 2025) to mask-free 3D editing. Vinedresser3D (Chi et al., 2026) automates mask generation with VLMs. Data-driven methods (Qi et al., 2024; Ma et al., 2025; Gat et al., 2026; Xu et al., 2026) instead require diverse paired edits with consistent non-edited regions. 3DEditVerse (Xia et al., 2025) provides approximately 116k training pairs through image editing and 3D reconstruction, whereas Nano3D (Ye et al., 2026b) uses native 3D editing to obtain a similarly sized dataset. Separate source and target reconstruction can compromise consistency. EditVerse3D (Yin et al., 2026) and PartFlow (Weng et al., 2026) improve consistency through part-aware operations, but covering diverse, imaginative instructions remains challenging. We introduce 3DEdit-1M for such instructions and the artist-modeled Consis3D-Bench to evaluate edit accuracy and consistency.

2.3 3D Unified Multimodal Model

Unified 3D modeling remains less explored than unified image understanding and generation (Deng et al., 2025; Xie et al., 2025b; Wang et al., 2024; Chen et al., 2025b). ShapeLLM-Omni (Ye et al., 2025) unifies understanding and generation autoregressively, while LUCID-3D (Anonymous, 2025) combines autoregressive and diffusion modeling. Cube3D (Bhat et al., 2025) uses self-supervised tokenization and supports scene generation. EVA01 (Yang et al., 2026) and Hunyuan3D-Buffalo (Ye et al., 2026a) couple understanding and flow-based generation experts. PriorEdit3D (Wen et al., 2026) distills 2D editing priors with 3D distribution matching. Scarce text–3D pairs still hinder knowledge transfer from pretrained VLMs. Our interleaved text–image–3D training enables native joint image and 3D generation, transferring visual priors to geometry for state-of-the-art text-to-3D performance.

Table 1: Training data composition. Tokenization configurations are provided in Supp. Tab. S2

Stage	Data Configuration	Items	Tokens/ Sample	Total Tokens	Token %
Pre-Training	Text + Image	63.7M	1,101	70.1B	14.8%
	Image + 3D	120M	2,048	245.8B	51.8%
	Text + 3D	9M	1,101	9.9B	2.1%
	Subtotal	*192.7M*		*— 325.8B*	*68.6%*
CT & SFT	Text + Single-View Image + 3D	2M	2,125	4.3B	0.9%
	Text + 6 Posed Images + 3D	20M	7,245	144.9B	30.5%
	Subtotal	*22M*		*— 149.2B*	*31.4%*
	Total	**214.7M**		**— 475.0B**	**100%**

3 Data Pre-processing

3.1 Data Overview

Our training corpus spans text, images, and 3D assets, organized into paired and triplet subsets. Tab. 1 summarizes data composition across pre-training and supervised fine-tuning stages. Token counts follow the per-modality configurations in Supp. Tab. S2. Detailed data synthesis, filtering, and captioning procedures are provided in Supp. D.

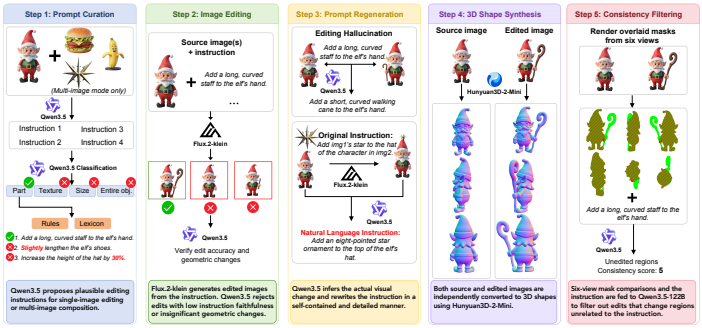

Figure 4: Overview of the 3D editing data synthesis pipeline.

Table 2: Pre-training data mixture.

Data Pool	Size	Tasks	Priority	Eff. P_i
Text–Image	63.7 M	Text→Image	1.0	0.22
Image–3D	120 M	Image→3D 3D→Image	1.5	0.40
Text–3D	9 M	Text→3D	3.0	0.37

3.2 3D Shape Editing Triplets

Real-world 3D editing data are scarce and costly to collect at scale. As discussed in Section 2.2, part-based data synthesis constrains the granularity of editing instructions. Other approaches prompt a VLM to imagine plausible modifications from a single image, often yielding commonplace, repetitive edits tied to object categories, such as modifying car bumpers or spoilers, human hairstyles, and animal horns or tails. We therefore introduce an additional multi-image composition pipeline to generate more diverse, open-ended instructions that better reflect real users' creative ideas. As shown in Fig. 4, a VLM selects a primary subject from two to four images and proposes edits to combine it with concepts or parts from the others. These instructions are fed into an image generation model, and the synthesized and original images are lifted to 3D. Finally, a VLM filters samples using multi-view renderings, rejecting 28.6% for unfaithful edits and 47.1% for inconsistencies in non-edited geometry. The resulting 3DEdit-1M contains 1M instruction–image–3D shape triplets, with 76.4% generated via multi-image composition and the rest via single-image editing under the same filtering strategy. Details and examples are provided in Supp. D.5.

Precise target shapes preserving non-edited regions are essential for evaluation, which can't be guaranteed by the synthetic data of existing benchmarks. Therefore, we reserve 450 source shapes and instructions from the 3DEdit-1M. Artists model their target shapes in ZBrush, strictly preserving non-target regions, forming Consis3D-Bench.

4 Model Training

4.1 Architecture

A key obstacle to unified multi-modal generation is the heterogeneity of representations across modalities. Text forms discrete symbol sequences, whereas images are typically encoded as 2D grids of continuous features and 3D shapes as volumetric lattices (Xiang et al., 2025b), unordered point sets, or continuous latent vectors (Zhang et al., 2023). Unifying these representations under a discrete autoregressive next-token objective poses two challenges: continuous latents require discretization, while spatial token counts grow with resolution, increasing attention cost; modality-specific positional encodings and attention patterns further complicate parameter sharing and cross-modal knowledge transfer. We address both challenges by converting continuous representations into discrete *1D* tokens, compressing each modality into a short, structurally uniform sequence. Omni123 adopts a self-developed representative image tokenizer (Section 4.2) for images and the Cube3D tokenizer (Bhat et al., 2025) for 3D meshes, enabling seamless concatenation without modality-specific positional encoding. Both modalities are then processed by a shared autoregressive Transformer backbone following the Cube3D architecture that maximizes parameter sharing and enforces cross-modal coherence through

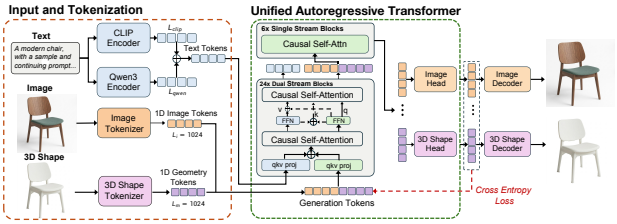

Figure 5: **Omni123**: a shared, text-conditioned autoregressive transformer for 1D image and geometry tokens.

shared self-attention, shown in Fig. 5. Detailed architecture design and hyperparameters are provided in Supp. A.

4.2 Image Tokenizer

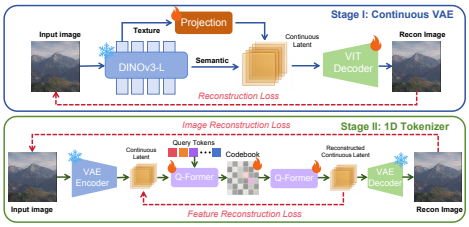

Figure 6: Two-stage image tokenizer training strategy.

Discrete tokenizers in auto-regressive models usually suffer from disjoint codebook optimization (Zhu et al., 2025), degrading reconstruction fidelity relative to continuous VAEs, limiting generation quality. To overcome this representation and optimization problem, we introduce a two-stage strategy (Fig. 6): (1) *Continuous VAE training.* We first learn high-fidelity continuous representations. Inspired by representation autoencoders (Zheng et al., 2025; Du et al., 2025; Shi et al., 2025), we retrain DINO-Tok (Jia et al., 2025) with a DINOv3-L encoder (Siméoni et al., 2025) on ImageNet (Deng et al., 2009). (2) *1D tokenizer training.* We freeze the Stage-I VAE and insert a trainable 1D tokenizer only to reconstruct continuous latents. This shifts vector quantization from direct pixel-space supervision to a simpler reconstruction task in a regularized, well-structured latent space. Training details are provided in Supp. B.5.

4.3 Model Pre-training (PT)

The central challenge motivating our pre-training design is the acute scarcity of 3D data relative to 2D: our corpus contains 63.7 M text–image pairs but only 9 M text–3D pairs. Naïvely mixing appearance-focused and geometry-focused objectives risks task interference rather than beneficial transfer. We address this challenge with the *X-to-X* paradigm: all cross-modal generation tasks—text-to-image, text-to-3D, image-to-3D, and 3D-to-image—are unified as autoregressive next-token prediction over a shared discrete vocabulary. By training a single model to traverse multiple generation paths across the "semantic–visual–geometric" cycle, the framework enforces cross-modal generative consistency as an implicit structural constraint. Concretely, the large-scale Text→Image subset strengthens the shared text conditioning pathway that Text→3D relies on despite its limited data, while the 3D→Image task—as the inverse of Image→3D—forces the model to maintain explicit 3D-to-2D geometric correspondences, which in turn benefits forward Image→3D reconstruction.

Training We train all four tasks with autoregressive next-token prediction. For the three data pools in Tab. 2, the Image–3D task direction is randomly selected per sample. We use temperature-based, priority-weighted sampling to balance the training data distribution. The sampling probability is $P_i = \frac{(\text{Size}_i)^T \times \text{Priority}_i}{\sum_{j=1}^{n} \left((\text{Size}_j)^T \times \text{Priority}_j\right)}$, where Size_i is the number of samples in pool i, Priority_i is a manually assigned importance weight reflecting task value, and $T \in (0, 1]$ is a temperature controlling the degree of rebalancing—lower T flattens the size distribution, giving smaller but high-priority pools greater representation. Detailed training setups are provided in Supp. B.1.

4.4 Continued Training (CT)

Pre-training establishes cross-modal alignment, yet the model lacks an explicit notion of *viewpoint*: it can generate images and 3D shapes, but cannot reason about how the same object appears from different camera poses.

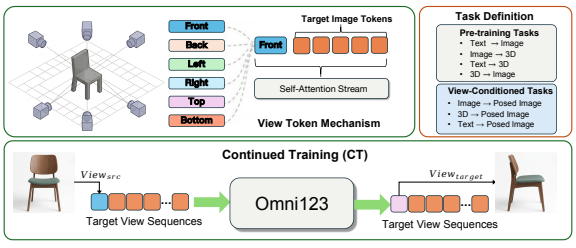

Figure 7: Overview of continued training.

This is a critical gap—multiple views of the same object share an underlying 3D structure. Predicting a novel view from an image or 3D geometry encourages the model to relate changes in appearance to this shared geometry. Continued training (CT) retains all four pre-training tasks and closes this gap by introducing three new view-conditioned tasks (Fig. 7). Specifically, we introduce

view tokens (Fig. 7): N learnable embeddings tied to fixed canonical viewpoints (front, back, left, right, top, bottom). Prepending a view token to the target image sequence enables viewpoint-controlled generation grounded in the shared 3D representation learned during pre-training. For image inputs, an additional source view token $\mathbf{v}_{\text{src}}$ specifies the input viewpoint, providing the model with explicit source–target viewpoint correspondence and enabling novel view synthesis.

Training setup We construct posed image–3D tuples by rendering each asset in the existing Image–3D data at six canonical viewpoints. During training, a task and a target viewpoint are randomly selected. The pre-training data pools remain active to prevent catastrophic forgetting of the base generation capabilities and new multi-view tuples are mixed with them using the same temperature-based, priority-weighted sampling strategy. Detailed data construction and training settings are provided in Supp. B.2.

4.5 Supervised Fine-Tuning (SFT)

PT and CT equip the model with cross-modal alignment and viewpoint awareness, but each task is still trained independently: there is no guarantee that Image→3D output renders back to the conditioning image or that Text→Image output yields a plausible 3D reconstruction.

The SFT stage closes the loop by introducing *interleaved multi-modal sequences* (Tab. 3) that chain image generation, 3D reconstruction, and novel view rendering into a single autoregressive sequence. Because each subsequent modality is conditioned on all previously generated tokens, the training signal naturally enforces cross-modal consistency: a 3D mesh must be faithful to the images that precede it, and novel views rendered from that mesh must be coherent with the original generation. Detailed data construction and training settings are provided in Supp. B.3.

Table 3: SFT task configurations

Task	Sequence
Text → Img → 3D	$[\text{Img}_1, \text{3D}]$
Text → 3D → Img	$[\text{3D}, \mathbf{v}_1, \text{Img}_1]$
$\text{Img}_{\times N}$ → 3D	$[\mathbf{v}_1, \text{Img}_1, \ldots, \mathbf{v}_N, \text{Img}_N, \text{3D}]$
3D → $\text{Img}_{\times N}$	$[\text{3D}, \mathbf{v}_1, \text{Img}_1, \ldots, \mathbf{v}_N, \text{Img}_N]$
Text → $\text{Img}_{\times N}$ → 3D → $\text{Img}_{\times M}$	$[\mathbf{v}_1, \text{Img}_1, \ldots, \mathbf{v}_N, \text{Img}_N, \text{3D}, \mathbf{v}_{N+1}, \text{Img}_{N+1}, \ldots, \mathbf{v}_{N+M}, \text{Img}_{N+M}]$

4.6 Instruction-Based 3D Editing

Earlier training stages equip Omni123 with strong 3D generation and cross-modal understanding capabilities by integrating text conditioning with rich 3D–2D correspondences. To adapt these capabilities to instruction-based 3D editing for practical content-creation workflows, we fine-tune the SFT model on 3DEdit-1M using the same autoregressive objective over target 3D tokens. Each training sample is organized as $[\text{Img}_{\text{src}}, \text{3D}_{\text{src}}, \text{3D}_{\text{tgt}}]$ and conditioned on text instructions. Detailed training settings are provided in Supp. B.4.

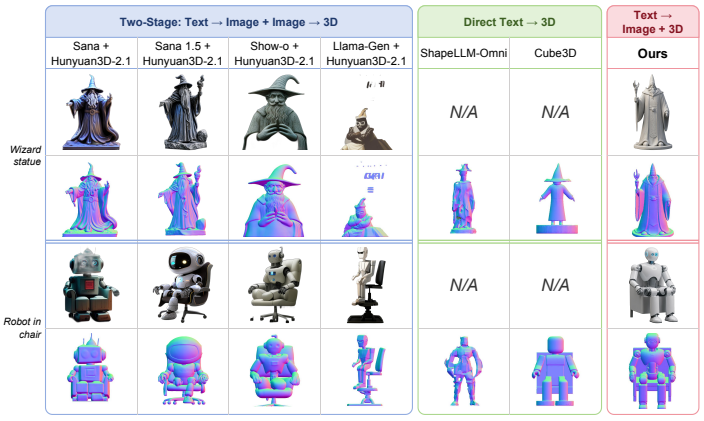

Figure 8: Qualitative comparison of text-to-3D. N/A denotes no image outputs.

Table 4: Quantitative performance of text-to-3D

Method	Param.	ULIP2-T ↑	Uni3D-T ↑
Two-stage: Text → Image → 3D (Hunyuan3D 2.1)			
Sana-1.5	4.8B	0.1516	0.1672
Sana	1.6B	0.1448	0.1589
Show-o	1.3B	0.1447	0.1581
LlamaGen	0.775B	0.0946	0.1067
Native: Text → 3D			
ShapeLLM-Omni	7B	0.1649	0.2666
Cube3D	1.6B	<u>0.1726</u>	<u>0.2768</u>
Ours	2.2B	**0.1829**	**0.2908**

5 EXPERIMENTS

In this section, we evaluate Omni123 on text-to-3D shape generation (Section 5.1), instruction-based 3D editing (Section 5.2), and image tokenizer reconstruction quality (Section 5.3). We provide more detailed comparisons on other X-to-X tasks in Supp. C.3. Unless otherwise stated, all Omni123 results use greedy decoding with Classifier-Free Guidance scale $\omega = 5.0$.

5.1 TEXT-TO-3D SHAPE GENERATION

As a cornerstone of 3D content generation, structurally consistent and high-quality shape generation provides the essential geometric scaffolding required for subsequent pipelines, including texturing, rendering, and physics simulation. Here, we evaluate the text-to-shape generation performance of Omni123, focusing on geometric fidelity and semantic alignment with input prompts.

Experiments To rigorously isolate the advantages of our **unified X-to-X training paradigm**, we benchmark against two dominant methodologies. ***(i) Cascaded text-to-image-to-3D pipelines:*** We pair leading image generators of *comparable parameter scales*—Sana-1.5 (4.8B), Sana (1.6B), Show-o (1.3B), and LlamaGen (0.775B)—with Hunyuan3D-2.1. ***(ii) Native text-to-3D models:*** ShapeLLM-Omni, a 7B autoregressive model that extends a VLM (Qwen2.5-VL) with 3D generation capabilities, and Cube3D, a 1.6B autoregressive model trained from scratch for text-to-3D generation. To evaluate semantic alignment, we uniformly sample $8{,}192$ surface points from each generated mesh and compute point-text similarities using ULIP-2 (Xue et al., 2024) and Uni3D (Zhou et al., 2023). We evaluate on 100 diverse text prompts encompassing various object categories and styles, compositional descriptions, and fine-grained attributes.

Comparisons Tab. 4 demonstrates that two-stage pipelines consistently underperform native methods in **semantic-geometric alignment**, primarily bottlenecked simultaneously by the inherent prior gap between the image generator and the 3D lifter and inevitable error accumulation across the cascaded pipeline. Native models avoid this handoff but previous works remain limited by scarce 3D supervision, ineffective knowledge transfer, and suboptimal modality fusion. By leveraging abundant 2D observations as a unified geometric prior, Omni123 overcomes these limitations and achieves SOTA performance in text–3D alignment. Notably, our **2.2B model** outperforms the 7B ShapeLLM-Omni, demonstrating that the native multi-task training paradigm is a fundamentally more *parameter-efficient*, scalable route than blindly scaling isolated generators upon a pretrained VLM backbone. Fig. 8 visually supports our quantitative findings on complex compositional prompts. More results are provided in Supp. E.2.

5.2 INSTRUCTION-BASED 3D EDITING

Instruction-based 3D editing enables users to adapt existing assets to new design requirements through natural-language instructions, which is essential for 3D content creation workflows. A successful edit must faithfully implement the requested change while preserving non-edited geometry. We evaluate these capabilities of Omni123 on Consis3D-Bench, Edit3D-Bench (Ma et al., 2025) (texture-related cases excluded), and 3DEditVerse (Xia et al., 2025).

Metrics To evaluate consistency in non-edited regions, *Consistency Distance* measures the nearest-neighbor distance from the ground-truth unchanged region to the generated mesh: $Dist_{\text{cons}} = \frac{1}{|\mathcal{P}_{\text{irr}}|} \sum_{\mathbf{p} \in \mathcal{P}_{\text{irr}}} \min_{\hat{\mathbf{p}} \in \hat{\mathcal{P}}} \|\mathbf{p} - \hat{\mathbf{p}}\|_2^2$, where $\mathcal{P}_{\text{irr}}$ denotes points sampled from the non-edited region, and $\hat{\mathcal{P}}$ denotes points sampled from the generated mesh surface.

Table 5: 3D editing results.

Method	Cons. Dist. ↓	ULIP2 ↑	DINO-I ↑	CLIP-I ↑	SSIM ↑	LPIPS ↓
Edit3D-Bench						
3DEditFormer	–	0.7830	0.7117	0.9117	0.8691	0.1981
Tailor3D	–	0.3542	0.3526	0.8330	0.8150	0.3542
ShapeLLM-Omni	–	0.5005	0.5252	0.8691	0.8521	0.2654
Steer3D	–	0.7467	0.7148	0.9149	0.8778	0.1611
Omni123 (Ours)	–	**0.7853**	**0.7644**	**0.9245**	**0.8946**	**0.1346**
3DEditVerse						
3DEditFormer	–	**0.8502**	0.7766	0.9326	0.8969	0.1294
Tailor3D	–	0.3399	0.3759	0.8521	0.8211	0.2979
ShapeLLM-Omni	–	0.5137	0.5235	0.8770	0.8458	0.2641
Steer3D	–	0.5723	0.5558	0.8806	0.8390	0.2739
Omni123 (Ours)	–	0.8239	**0.7865**	**0.9330**	**0.9177**	**0.1060**
Consis3D-Bench						
3DEditFormer	0.2265	0.7450	0.6648	0.8961	0.8523	0.2087
Tailor3D	0.0733	0.6800	0.5462	0.8872	0.8459	0.1769
ShapeLLM-Omni	0.1573	0.5577	0.5535	0.8716	0.8338	0.2444
Steer3D	0.1828	0.4518	0.4874	0.8525	0.8141	0.3102
Omni123 (Ours)	**0.0134**	**0.9272**	**0.8859**	**0.9595**	**0.9296**	**0.0583**

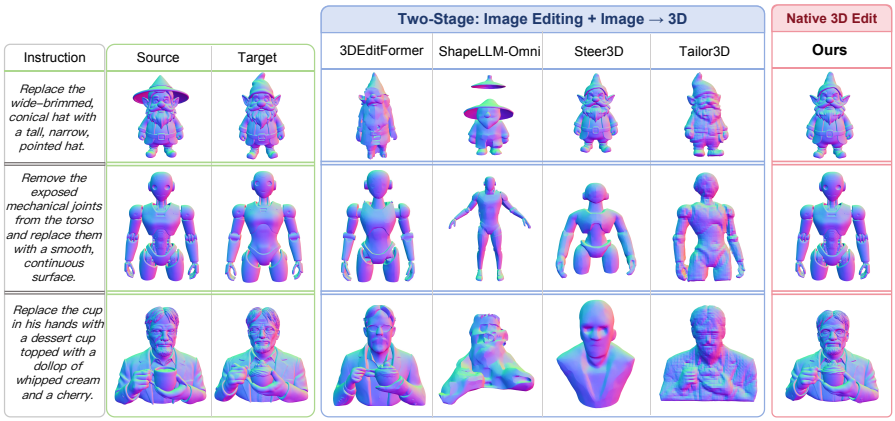

Figure 9: Instruction-based 3D editing. Omni123 follows edit instructions more faithfully and better preserves non-edited geometry than two-stage image-editing pipelines.

We report this metric only on Consis3D-Bench: the other benchmarks use shapes synthesized by image-to-3D models, making reliable ground-truth unchanged regions difficult to obtain. For edit accuracy, we measure cosine similarity between the ULIP-2 features of generated and ground-truth point clouds. Following Xia et al. (2025), we also compute DINO-I, CLIP-I, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) on multi-view normal maps to assess shape similarity.

Comparisons We compare against ShapeLLM-Omni, Steer3D (Ma et al., 2025), 3DEditFormer (Xia et al., 2025), and Tailor3D (Qi et al., 2024). Baseline inference details are provided in Supp. C.6. Tab. 5 and Fig. 9 show strong editing performance across the three benchmarks. On Consis3D-Bench, Omni123 achieves the best results across all six metrics. Without training on Edit3D-Bench or the 3DEditVerse training set, Omni123 also achieves competitive zero-shot performance on both benchmarks. These results support the benefits of our curated dataset and native multimodal interleaved training. Comparisons with additional 3D editing baselines are provided in Supp. C.6. We exclude these methods from the main results because they require masks or edited images beyond text instructions.

5.3 Image Tokenizer Evaluation

A high-fidelity image tokenizer dictates essential structural and textural constraints and mediates the cross-modal information flow. We evaluate our image tokenizer (Section 4.2) on the ImageNet-1K 512×512 test set. Following previous works, we report FID (Heusel et al., 2017), PSNR and SSIM.

Comparisons Tab. 6 shows that our VAE matches the powerful Flux.1 (Black Forest Labs, 2024) in FID and surpasses it in PSNR and SSIM, despite using only the ImageNet training set. This highlights the value of the pretrained DINOv3 (Siméoni et al., 2025) representation prior. Our discrete 1D pipeline also achieves the best scores across all metrics, supporting the two-stage strategy for limiting quantization degradation. Fig. 10 complements these results: previous methods (Agarwal

Table 6: Quantitative comparison of image tokenizers on ImageNet-1K 512×512 test set.

Method	FID ↓	PSNR ↑	SSIM ↑
Continuous			
Cosmos-CI16 × 16	0.356	28.60	0.783
Flux.1	**0.017**	37.69	0.963
SDXL	0.188	30.35	0.835
SVG-T2I	1.994	20.60	0.587
Ours VAE	**0.017**	**39.15**	**0.966**
Discrete			
Cosmos-DI16 × 16	1.215	23.76	0.645
LlamaGen	0.839	23.96	0.664
Show-o	1.183	24.20	0.680
Ours VAE + 1D Tokenizer	**0.432**	**24.61**	**0.759**

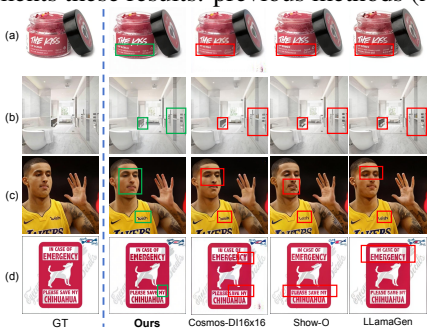

Figure 10: Qualitative comparison of discrete image tokenizers.

et al., 2025; Xie et al., 2024b; Sun et al., 2024) often introduce blur and semantic distortion, while our tokenizer better preserves high-frequency details and structural integrity. It retains faithful text and logos (Rows *a*, *c*, and *d*), sharp boundaries in complex scenes (Row *b*), and consistent facial features (Row *c*). These capabilities preserve the geometric and semantic cues needed to guide joint 3D synthesis. Supp. C.4 provides an ablation comparing our tokenizer with GiGaTok (Xiong et al., 2025b) to assess its contribution to generation quality and unified text–image–3D modeling.

6 Ablation Study

We examine the training paradigm and the performance gains from scaling 3D editing data in this section. Additional ablations of task composition and heterogeneous data composition are provided in Supp. C.2 and Supp. C.1, respectively, and a user study is provided in Supp. C.5.

6.1 Pretraining Strategies

Tab. 7 compares joint multi-task PT *(a)* with interleaved X-to-X PT *(b)* without SFT. Joint PT performs better on both metrics and serves as the baseline for the subsequent SFT comparisons. We next evaluate whether closing the image–3D–image cycle during SFT provides an implicit structural constraint on 3D learning. Starting from the same PT baseline (a), we compare a single-view forward chain (c) with the full multi-view cycle (d). Row (c) yields only a small gain, whereas Row (d) improves both metrics significantly. This improvement stems from additionally predicting posed views after the 3D segment, encouraging geometry that supports consistent renderings. With the same underlying input–target pairs, coupling these predictions in one sequence links the preceding and subsequent images through shared geometry.

Table 7: Ablation of training paradigm.

Configuration	ULIP2-T ↑	Uni3D-T ↑
pre-train		
(a) **Joint multi-task PT**	**0.1745**	**0.2775**
(b) Interleaved X-to-X PT	0.1659	0.2638
SFT from (a)		
(c) Single-view SFT	0.1755	0.2790
(d) **Multi-view cycle (Ours)**	**0.1829**	**0.2908**

6.2 Scaling with 3DEdit Data

We study how editing performance scales with supervision by varying the number of 3DEdit-1M training triplets from 100k to 1M and evaluating on Consis3D-Bench. Fig. 11 shows consistent gains in both geometry preservation and edit accuracy: Consistency Distance decreases, while ULIP-2 increases. Both metrics continue to improve from 100k to 1M, indicating no clear saturation within the tested range. We provide additional ablations in Supp. C.7 and qualitative examples in Figs. S1 and S2 to further verify our data quality.

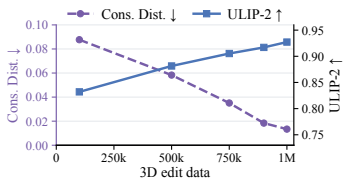

Figure 11: Scaling with 3D editing data.

7 Conclusion

We presented Omni123, a unified multimodal generative model that addresses 3D data scarcity by unifying text-to-2D and text-to-3D generation within a single autoregressive framework, enabling abundant 2D observations to serve as an implicit structural constraint for 3D learning. Experiments demonstrate that our unified approach significantly improves geometric fidelity and semantic alignment, offering a scalable pathway toward multimodal 3D world models. Current limitations include the fixed mesh resolution of the 3D tokenizer, the lack of texture generation, reliance on canonical viewpoints, and exact vertex-level consistency in unedited regions. Future work will explore adaptive-resolution tokenization, material modeling and unified image–3D editing.

AI USE STATEMENT

In this work, we used generative AI tools to synthesize, caption, and filter training data, as described in Section 3 and Supp. D, with details of the editing-data pipeline in Supp. D.5. We also used these tools to prepare auxiliary inputs for baseline evaluation (Supp. C.6). AI assistance with research code was limited to data visualization; it was not used to implement our models or training and evaluation pipelines. Each synthetic-data pipeline described in this paper was checked by at least five human reviewers, who examined its generated outputs and filtering decisions and confirmed their consistency with human judgment.

Additionally, we used AI tools for polish writing, correct grammatical errors, and reference formatting. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Anonymous. LUCID-3D: A lightweight and compatible framework for unified 3D understanding and generation. OpenReview preprint, 2025. URL https://openreview.net/forum?id=4mOyMTmTD4. Submitted to ICLR 2026.

Sangwon Baik, Hyeonwoo Kim, and Hanbyul Joo. Learning 3D object spatial relationships from pre-trained 2D diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8418–8428, 2025.

Kiran Bhat, Nishchaie Khanna, Karun Channa, Tinghui Zhou, Yiheng Zhu, Xiaoxia Sun, Charles Shang, Anirudh Sudarshan, Maurice Chu, Daiqing Li, et al. Cube: A Roblox view of 3D intelligence. *arXiv preprint arXiv:2503.15475*, 2025.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Yankuan Chi, Xiang Li, Zixuan Huang, and James M. Rehg. Vinedresser3D: Towards agentic text-guided 3D editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2602.19542.

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. URL https://arxiv.org/abs/2505.14683.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Sinan Du, Jiahao Guo, Bo Li, Shuhao Cui, Zhengzhuo Xu, Yifu Luo, Yongxian Wei, Kun Gai, Xinggang Wang, Kai Wu, et al. VQRAE: Representation quantization autoencoders for multimodal understanding, generation and reconstruction. *arXiv preprint arXiv:2511.23386*, 2025.

Inbar Gat, Dana Cohen-Bar, Guy Levy, Elad Richardson, and Daniel Cohen-Or. ShapeUP: Scalable image-conditioned 3D editing. In *SIGGRAPH 2026 Conference Papers*. ACM, 2026. doi: 10.1145/3799902.3811073. URL https://doi.org/10.1145/3799902.3811073.

Google. Nano Banana 2 – Gemini AI image generator & photo editor, 2026. URL https://gemini.google/overview/image-generation/.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Hunyuan3D Team, 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.

Mingkai Jia, Mingxiao Li, Liaoyuan Fan, Tianxing Shi, Jiaxin Guo, Zeming Li, Xiaoyang Guo, Xiao-Xiao Long, Qian Zhang, Ping Tan, et al. DINO-Tok: Adapting DINO for visual tokenizers. *arXiv preprint arXiv:2511.20565*, 2025.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. FlowEdit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025. URL https://arxiv.org/abs/2412.08629.

Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. GaussianAnything: Interactive point cloud flow matching for 3D generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P4DbTSDQFu.

Lin Li, Zehuan Huang, Haoran Feng, Gengxiong Zhuang, Rui Chen, Chunchao Guo, and Lu Sheng. VoxHammer: Training-free precise and coherent 3D editing in native 3D space. *arXiv preprint arXiv:2508.19247*, 2025a. URL https://arxiv.org/abs/2508.19247.

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

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Zhen Liu, Yao Feng, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. MeshDiffusion: Score-based generative 3D mesh modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0cpM2ApF9p6.

Ziqi Ma, Hongqiao Chen, Yisong Yue, and Georgia Gkioxari. Feedforward 3D editing via text-steerable image-to-3D. *arXiv preprint arXiv:2512.13678*, 2025.

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. URL https://arxiv.org/abs/2212.08751.

Rui Nie, Chuang Wang, Haitao Zhou, Jiahe Song, Buyu Li, Sheng Wang, and Qian Yu. EditFlow3D: Automated Local Editing of 3D Assets with Trajectory Preservation. *arXiv preprint arXiv:2608.03179*, 2026. URL https://arxiv.org/abs/2608.03179.

OpenAI. Introducing 4o image generation, 2025. URL https://openai.com/index/introducing-4o-image-generation/.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Zhangyang Qi, Yunhan Yang, Mengchen Zhang, Long Xing, Xiaoyang Wu, Tong Wu, Dahua Lin, Xihui Liu, Jiaqi Wang, and Hengshuang Zhao. Tailor3D: Customized 3D assets editing and generation with dual-side images. *arXiv preprint arXiv:2407.06191*, 2024.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.

Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. RichDreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3D. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9914–9925, 2024.

Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-E: Text-guided voxel editing of 3D objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 430–440, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Sella_Vox-E_Text-Guided_Voxel_Editing_of_3D_Objects_ICCV_2023_paper.html.

Etai Sella, Hao Phung, Nitay Amiel, Or Litany, Or Patashnik, and Hadar Averbuch-Elor. Prox-E: Fine-grained 3D shape editing via primitive-based abstractions. *arXiv preprint arXiv:2604.23774*, 2026. URL https://arxiv.org/abs/2604.23774.

Minglei Shi, Haolin Wang, Borui Zhang, Wenzhao Zheng, Bohan Zeng, Ziyang Yuan, Xiaoshi Wu, Yuanxing Zhang, Huan Yang, Xintao Wang, et al. SVG-T2I: Scaling up text-to-image latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2512.11749*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Haoyuan Wang, Zhenwei Wang, Xiaoxiao Long, Cheng Lin, Gerhard Hancke, and Rynson WH Lau. MAGE: Single image to material-aware 3D via the multi-view G-Buffer estimation model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10985–10995, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Hao Wen, Weibin Yun, Hongxing Fan, Haotian Lu, Rui Chen, Zehuan Huang, and Lu Sheng. Learning 3D editing without paired supervision via generative prior distillation. In *SIGGRAPH Asia 2026 Conference Papers*, SA Conference Papers '26, 2026. doi: 10.1145/3829340.3842352. URL https://doi.org/10.1145/3829340.3842352.

Jiawei Weng, Saining Zhang, Zhenxin Diao, Peishuo Li, Henghaofan Zhang, Junhao Chen, and Hao Zhao. Feedforward 3D editing learns from semantic-part transformation. *arXiv preprint arXiv:2605.27351*, 2026. URL https://arxiv.org/abs/2605.27351. Accepted to SIGGRAPH Asia 2026.

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. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.14832.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Philip Torr, Xun Cao, and Yao Yao. Direct3D-S2: Gigascale 3D generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412*, 2025. URL https://arxiv.org/abs/2505.17412.

Ruihao Xia, Yang Tang, and Pan Zhou. Towards scalable and consistent 3D editing. *arXiv preprint arXiv:2510.02994*, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3D generation. *arXiv preprint arXiv:2512.14692*, 2025a.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21469–21480, 2025b.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024b.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025b.

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*, 2025a. doi: 10.1111/cgf.70198. URL https://arxiv.org/abs/2408.14732.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. GigaTok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18770–18780, 2025b.

Yizhao Xu, Hongyuan Zhu, Caiyun Liu, Tianfu Wang, Keyu Chen, Sicheng Xu, Jiaolong Yang, Nicholas Jing Yuan, and Qi Zhang. Beyond Voxel 3D Editing: Learning from 3D Masks and Self-Constructed Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 2026. URL https://arxiv.org/abs/2604.13688.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP-2: Towards scalable multimodal pre-training for 3D understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2305.08275.

Zongyuan Yang, Mingjing Yi, Wanli Ma, Chenzhuo Fan, Bocheng Li, Baolin Liu, Yuke Lou, Yingde Song, Yongping Xiong, Zhengdong Guo, and Shimu Wang. EVA01: Unified Native 3D Understanding and Generation via Mixture-of-Transformers. *ACM Transactions on Graphics*, 2026. URL https://arxiv.org/abs/2605.16745. Accepted; SIGGRAPH Asia 2026.

Junliang Ye, Zhengyi Wang, Ruowen Zhao, Shenghao Xie, and Jun Zhu. ShapeLLM-Omni: A native multimodal LLM for 3D generation and understanding. *arXiv preprint arXiv:2506.01853*, 2025.

Junliang Ye, Kenkun Liu, Guocun Wang, Yang Li, Yansong Qu, Chunshi Wang, Jingwei Xu, Yunhan Yang, Zibo Zhao, Jiachen Xu, Jiaao Yu, Lifu Wang, Zhihao Liang, Xin Huang, Zhuo Chen, and Chunchao Guo. Hunyuan3D-Buffalo 1.0: A Unified Multimodal Model for Scalable 3D Generation, Understanding, and Editing. *arXiv preprint arXiv:2608.02711*, 2026a. URL https://arxiv.org/abs/2608.02711.

Junliang Ye, Shenghao Xie, Ruowen Zhao, Zhengyi Wang, Hongyu Yan, Wenqiang Zu, Lei Ma, and Jun Zhu. Nano3D: A training-free approach for efficient 3D editing without masks. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2510.15019.

Youtan Yin, Yanning Zhou, Jiacheng Wei, Xiaofeng Yang, Jun Zhang, Jiayang Bai, Jingwen Ye, Weidong Zhang, and Guosheng Lin. EditVerse3D: High-Quality 3D Object Editing with Region-Aware Learning. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2607.07187. To appear.

Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3DShape2VecSet: A 3D shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. CLAY: A controllable large-scale generative model for creating high-quality 3D assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1801.03924.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

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.

Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22968–22977, 2025.