# CoRe3D: Coupling Semantic Reasoning and Spatial Construction for Unified 3D Generation and Understanding

**Anonymous authors**
Paper under double-blind review

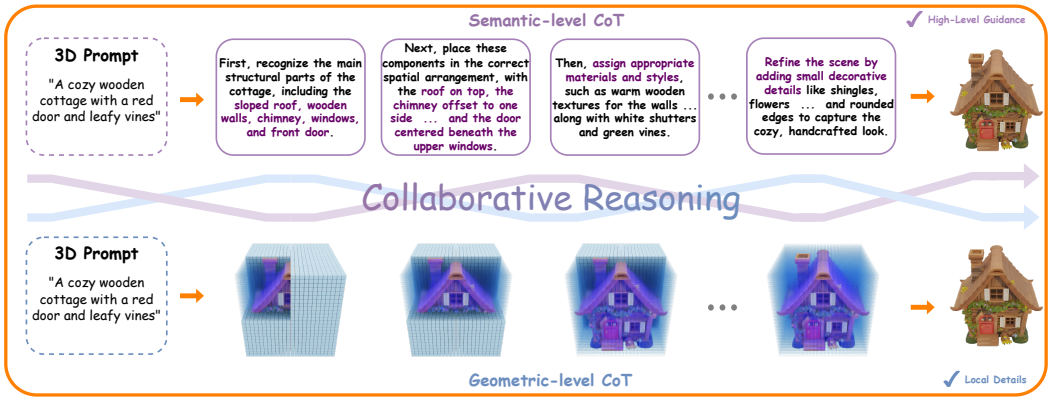



Figure 1: We introduce **CoRe3D**, a unified framework for 3D understanding and generation that couples Semantic CoT reasoning with octant-based Geometric CoT (geometric construction trace) through collaborative reasoning. Through semantic–geometric co-attention, CoRe3D uses textual plans to guide 3D construction and observed geometry to ground textual descriptions of objects.

## Abstract

Recent progress in multimodal foundation models shows that explicit reasoning mechanisms can improve reliability, cross-modal alignment, and generation, yet comparable mechanisms for 3D remain limited. We present **CoRe3D**, a unified framework for 3D understanding and generation, that performs collaborative reasoning over two coupled traces: (i) a semantic CoT that expands a prompt into an interpretable plan (parts, attributes, relations), and (ii) an octant-based geometric CoT that realizes this plan as localized 3D tokens, enabling compositional and interpretable generation. To jointly optimize both traces end-to-end under heterogeneous, partly non-differentiable 3D objectives, we propose **Co-GRPO**, a trace-aware group-relative preference optimization that integrates multi-critic 3D rewards. Experiments show that CoRe3D improves 3D understanding (*e.g.*, +27% BLEU-1, +23% ROUGE-L, +13% METEOR over the strongest baseline) and achieves state-of-the-art text-to-3D alignment (+38% CLIP over the next best method), while also supporting a diverse set of tasks spanning general multimodal reasoning, image-to-3D, reasoning-based 3D generation over complex prompts, and fine-grained 3D part editing within a single model.

## 1 Introduction

Despite rapid progress in 3D generation, most methods remain reconstruction-driven, optimizing a generative model to match observed 3D shape or appearance signals (*e.g.*, rendered views or distillation targets) (Poole et al., 2022; Yi et al., 2024) rather than explicitly reasoning object-level structure and intent. As a result, they often struggle with prompts that are indirect or underspecified, a limitation that directly hinders generalization to open-ended, long-tail requests. In contrast, recent unified vision–language models exhibit reliable reasoning in 2D settings (Yang et al., 2024a; Touvron et al., 2023). A key contributor to this progress is the integration of Chain-of-Thought (CoT) reasoning (Wei et al., 2022), which, when extended to multimodal LLMs (Chen et al., 2025a;



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.

Meng et al., 2025; Zhang et al., 2025; 2024c), improves interpretability and consistency across visual reasoning tasks (Lu et al., 2024; Jiang et al., 2025). However, analogous unified reasoning in 3D remains underexplored. While a growing number of vision–language models support 3D understanding (Cheng et al., 2024; Shen et al., 2025; Chen et al., 2024b; Peng et al., 2023; Liu et al., 2023b; Huang et al., 2025a), only a handful unify understanding and generation in a single framework (Ye et al., 2025b; Wang et al., 2024b), let alone optimize the CoT reasoning during training.

To bridge this gap, we propose **CoRe3D**, a 3D Unified Language Model (3D-ULM) that brings *optimizable reasoning* to both 3D generation and 3D understanding within a single backbone. Our central idea is to couple explicit semantic reasoning with spatially structured 3D construction (Fig. 1). The model first generates a semantic chain of thought (Semantic CoT) that describes the intended object, its attributes, and spatial relations. It then constructs the object as a sequence of localized octant codes conditioned on this description (Geometric CoT). A semantic–geometric **co-attention interface** gives each geometric prediction direct access to the semantic trace, allowing the textual plan to guide construction throughout decoding. To train, we introduce **Co-GRPO**, a group-relative preference objective that jointly optimizes *generation* and *understanding* rollouts using trace-aware multi-critic 3D rewards. In summary, our contributions are:

- We introduce **CoRe3D**, a unified framework that couples semantic CoT reasoning with an octant-based geometric construction trace. This formulation connects explicit textual descriptions to spatially indexed 3D generation without requiring category-specific or part-specific ontologies.
- To jointly optimize semantic and geometric reasoning in 3D, we propose **Co-GRPO**, a group-relative preference objective that trains both *generation* and *understanding* rollouts with trace-aware multi-critic 3D rewards and adaptive balancing under heterogeneous supervision. Together with the **co-attention interface**, this is the first framework to optimize coupled semantic-geometric reasoning traces in a unified 3D model.
- We show that **CoRe3D** enables *reasoning-intensive text-to-3D generation* beyond literal prompt matching. On standard Text-to-3D benchmarks, it delivers strong quantitative gains over the best baseline (CLIP 0.29→0.40, ↑37.9%, with improved FD/KD), and qualitatively succeeds on indirect, knowledge-heavy prompts that require semantic inference (*e.g.*, generating the Statue of Liberty from descriptive cues without naming it), producing shapes faithful to the intended concept.
- We further demonstrate that our unified reasoning paradigm is not limited to generation but naturally improves reciprocal 3D understanding tasks, such as 3D-to-text captioning, highlighting its potential as a scalable foundation for general 3D intelligence.

## 2 Related Work

**3D Generation.** Early text-to-3D frameworks formulated 3D synthesis as an optimization problem guided by 2D priors through score distillation sampling (SDS) (Poole et al., 2022; Wang et al., 2023c; Chen et al., 2023; Lin et al., 2023; Raj et al., 2023; Li et al., 2023; Sun et al., 2023; Chen et al., 2024g; Wang et al., 2022; Tang et al., 2023; Yi et al., 2024). While this approach enabled cross-modal 3D generation without paired data, it required long per-instance optimization and often produced view-inconsistent geometry. Subsequent methods (Wang & Shi, 2023; Shi et al., 2023b; Wang et al., 2023b; Ye et al., 2024b; Qiu et al., 2024; Chen et al., 2025b) address view inconsistency by enforcing cross-view semantic constraints within diffusion pipelines. Other works address the inefficiency of iterative optimization by first generating consistent 2D renderings, then reconstructing the 3D geometry through fast neural reconstruction (Long et al., 2024; Zhao et al., 2024; Liu et al., 2023e;d; Shi et al., 2023a; Weng et al., 2023; Liu et al., 2023c; Wu et al., 2024a; Chen et al., 2024h; Voleti et al., 2024; Ye et al., 2024a; Liu et al., 2024a; Yu et al., 2025; Gao et al., 2024).

More recently, native 3D diffusion models shifted toward generative modeling within latent 3D spaces, employing VAE-based encoders to learn volumetric or implicit shape priors (Zhao et al., 2023; Wang et al., 2023a; Wu et al., 2024b; Yang et al., 2024b; Huang et al., 2025c; Zhang et al., 2024a; Xiang et al., 2024; Chen et al., 2024e; Li et al., 2024; Ye et al., 2025a). In contrast to diffusion models, approaches employ vector-quantized autoencoders (Van Den Oord et al., 2017), casting 3D generation as an autoregressive sequence modeling problem (Siddiqui et al., 2024; Chen et al., 2024c; Weng et al., 2024a). Later works introduce task-specific tokenization schemes that directly encode vertex–face structures, improving geometric fidelity and local continuity (Chen et al., 2024d;a; Weng et al., 2024b; Tang et al., 2024; Hao et al., 2024; Zhao et al., 2025; Wang et al.,

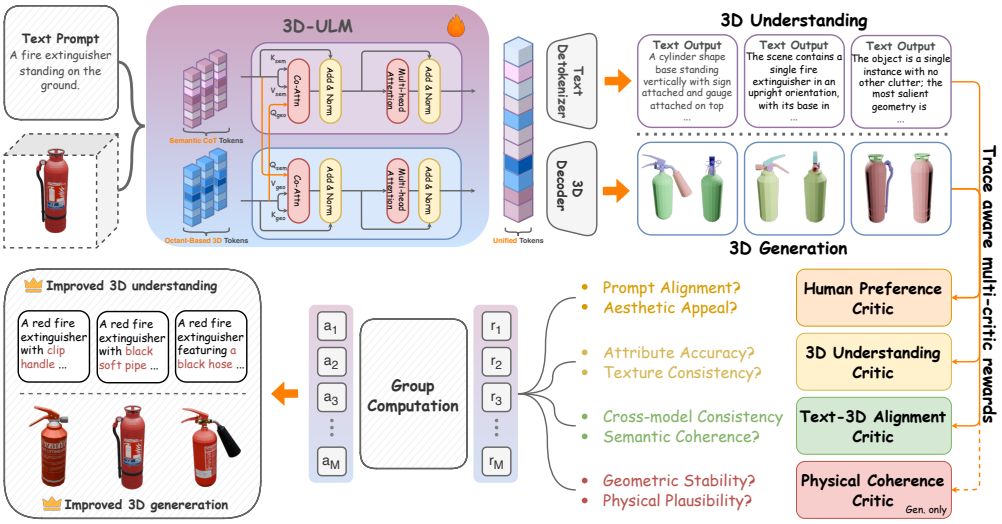



Figure 2: **CoRe3D overview.** Given a prompt and/or an observed 3D input, our 3D-ULM produces coupled semantic and geometric reasoning traces via semantic-geometric co-attention. Each rollout is rendered and evaluated by trace-aware multi-critic rewards. Co-GRPO performs group-relative preference optimization by computing within-group advantages and updating the unified 3D-ULM.

2024b). Building on autoregressive 3D generation, CoRe3D introduces an explicit semantic trace and a co-attention interface that conditions spatial token prediction on this trace. The semantic and geometric outputs are jointly optimized using 3D-aware rewards, connecting the model's textual description of the intended object with its geometric construction.

**Unified Generation & Understanding.** Recent multimodal LLMs have revealed remarkable capability in jointly processing and generating multimodal content. Early frameworks (Alayrac et al., 2022; Bai et al., 2023; Chen et al., 2024f) incorporate visual encoders for grounded perception, while more recent systems (Team, 2024; Liu et al., 2024b; Wang et al., 2024a; Xie et al., 2024; Zhou et al., 2024) integrate text and image generation through learned visual tokenizers and mixed-modality training. Extending this paradigm to 3D, emerging studies (Hong et al., 2023; Xu et al., 2024; Qi et al., 2024; Chen et al., 2024b) adapt LLMs for 3D understanding using point-cloud or shape embeddings. While effective for perception tasks, these models largely focus on recognition rather than generation. Subsequent efforts (Yin et al., 2025; Wang et al., 2024b; Chen et al., 2025c; Ye et al., 2025b; Zhu et al., 2023) attempt to unify language and 3D modeling by developing generative LLMs that handle both understanding and generation within a shared representation space. More interactive paradigms, such as LL3M (Lu et al., 2025) and L3GO (Yamada et al., 2025), employ agent-based reasoning to iteratively construct or edit 3D scenes, yet rely on symbolic planning rather than token-level 3D reasoning. In contrast, we introduce Co-GRPO, a collaborative group-relative preference optimization strategy that jointly evaluates semantic and geometric rollouts using complementary 3D-aware critics, enabling reasoning-aware 3D understanding and generation, and bridging the gap between linguistic intent and physically grounded 3D synthesis.

## 3 METHOD

A core challenge in unified 3D foundation models is translating open-ended language into structured reasoning signals that preserve compositional semantics and physical constraints. Directly mapping prompts to latent 3D tokens is inherently under-specified, as language descriptions typically omit precise geometric details, relations, and material cues, resulting in shapes that match coarse appearance but break in structure, texture consistency, or global coherence. We address this by optimizing 3D reasoning through two coupled traces (Section 3.1): (i) a semantic CoT that expands the prompt into detailed descriptions, and (ii) a geometric CoT that realizes the description as discrete octant tokens. To connect the two streams, we introduce a lightweight **Co-Attention Interface** (Section 3.2). Then, to optimize them, we introduce **Co-GRPO** (Section 3.3), a collaborative multi-

critic group-relative preference objective that encourages semantic faithfulness, text-3D alignment, and physical coherence (see Figure 2). We summarize the full training procedure in Section 3.4.

### 3.1 Semantic and Geometric Representations

**Semantic Representations.** Given a text prompt, CoRe3D first produces a detailed natural language description of the object category, spatial layout, materials, and appearance details, serving as an interpretable, text-based scaffold to anchor the subsequent geometric reasoning process. Formally, we represent the semantic CoT trace as a sequence of tokens $\mathcal{S}_{\text{sem}} = [s_1, s_2, \ldots, s_N]$, conditioned on the input prompt and optimized jointly with 3D generation tokens, enabling semantic intent to directly modulate spatial synthesis during training.

**Geometric Representations.** We represent each 3D object as a $64^3$ voxel grid and use a 3D VQ-VAE encoder to map it to a $16^3$ latent grid containing local geometry and appearance features. To obtain a compact sequence, we partition the latent grid into non-overlapping $2\times2\times2$ blocks and concatenate the feature vectors of the eight cells within each block. This reduces the 4096 latent vectors to 512 block features, each represented by a 64-dimensional vector. We refer to each such neighborhood as a local *octant block*. A vector-quantization module with an 8192-entry codebook maps each block feature to a discrete code, yielding 512 discrete 3D tokens per object. We serialize these tokens in Morton (Z-order) over the $8\times8\times8$ block grid. To distinguish spatial locations, we add a learned absolute position embedding to each quantized token embedding, indexed by the block's coordinates $(x_b, y_b, z_b)$. Position embeddings are added after quantization, allowing the codebook to encode local content while providing the generator with explicit spatial information. This representation defines a *geometric construction trace*, $\mathcal{G}_{\text{geo}} = [g_1, g_2, \ldots, g_M]$, where $M = 512$ and each $g_i$ specifies a codebook entry for a spatially indexed octant block. During generation, the model predicts these tokens autoregressively, conditioning each prediction on the input, the semantic reasoning trace, and previously generated geometric tokens. The semantic trace describes the intended object, its attributes, and spatial relations, while the geometric trace encodes its construction in the 3D latent space. Once the complete sequence is generated, the VQ-VAE decoder reconstructs the voxel grid.

### 3.2 Semantic-Geometric Co-Attention

In a unified 3D model, the semantic CoT should function as an explicit intermediate representation that supports both generation and understanding. However, if the semantic CoT is used only as a static prefix, its fine-grained constraints may be weakly utilized once long-horizon geometric decoding begins. To make semantic-geometric collaboration explicit and persistent, we introduce a lightweight *co-attention interface* to connect semantic and geometric tokens. Additionally, this co-attentive coupling defines the rollout interface optimized by Co-GRPO in Sec. 3.3.

More formally, let $H^{\text{sem}} \in \mathbb{R}^{N\times d}$ denote the hidden states of the semantic trace $\mathcal{S}_{\text{sem}}$, and let $h_t^{\text{geo}} \in \mathbb{R}^d$ denote the geometric hidden state used to predict the $t$-th octant token $g_t$. Before predicting octant token $g_t$, we update $h_t^{\text{geo}}$ by attending to the semantic trace:

$$\tilde{h}_t^{\text{geo}} = h_t^{\text{geo}} + \mathrm{CrossAttn}\big(h_t^{\text{geo}}, H^{\text{sem}}\big), \tag{1}$$

where $\mathrm{CrossAttn}(\cdot)$ is a standard multi-head cross-attention block. This ensures attributes and relations expressed in $\mathcal{S}_{\text{sem}}$ remain accessible throughout the geometric decoding process. In practice, for generation, geometric tokens are generated conditioned on $[T_p; \mathcal{S}_{\text{sem}}]$. For understanding, the 3D token stream is provided by encoding the observed input $V$ (voxel grid) to octant tokens $\mathcal{G}_{\text{geo}}^V$, and the same co-attention interface couples $\mathcal{S}_{\text{sem}}$ with $\mathcal{G}_{\text{geo}}^V$ during text decoding.

### 3.3 Collaborative Reasoning: Co-GRPO

A unified 3D foundation model must learn to both *generate* 3D assets from language and *understand* observed 3D inputs via text outputs, yet these two modes exhibit markedly different supervision and failure modes. Generation is constrained by non-differentiable criteria such as multi-view realism and physical coherence, while understanding is constrained by 3D-grounded semantic correctness. Bridging this gap with a single training objective is challenging because reward signals are heterogeneous, noisy, and often miscalibrated across prompts and modalities. We therefore propose **Co-GRPO**, a novel unified group-relative preference objective that learns from *relative compar-*

*isons within each input group*, yielding stable signals that align both generation and understanding under shared model parameters even when individual critics are imperfect.

Co-GRPO operates on two rollout types under a shared policy $\pi_\theta$. For *generation* rollouts, given a text prompt $T_p$ we sample a group of joint semantic-geometric rollouts, indexed by $i$:

$$(\mathcal{S}_{\text{sem}}^{(i)}, \mathcal{G}_{\text{geo}}^{(i)}) \sim \pi_\theta(\cdot \mid T_p), \quad i = 1, \ldots, G, \tag{2}$$

where each rollout first generates a semantic CoT $\mathcal{S}_{\text{sem}}^{(i)}$ and then generates geometric tokens $\mathcal{G}_{\text{geo}}^{(i)}$ conditioned on $[T_p; \mathcal{S}_{\text{sem}}^{(i)}]$. For *understanding* rollouts, given a text query and an observed voxel grid $(T_q, V)$, we first encode $V$ into octant-based 3D tokens $\mathcal{G}_{\text{geo}}^{V}$, then sample a text response conditioned on the query and the encoded tokens:

$$\mathcal{S}_{\text{sem}}^{(i)} \sim \pi_\theta\left(\cdot \mid T_q, \mathcal{G}_{\text{geo}}^{V}\right), \qquad o_{\text{text}}^{(i)} \sim \pi_\theta\left(\cdot \mid T_q, \mathcal{G}_{\text{geo}}^{V}, \mathcal{S}_{\text{sem}}^{(i)}\right), \qquad i = 1, \ldots, G. \tag{3}$$

In both cases, we treat each sampled output as a rollout and compute group-relative advantages within each input group for policy updates.

**Trace-aware Multi-critic Rewards.** A key design of Co-GRPO is *trace-aware* evaluation: rollouts are scored in the context of their own semantic trace $\mathcal{S}_{\text{sem}}^{(i)}$, enabling joint credit assignment. For generation, the semantic trace induces language-specific supervision signals in addition to conditioning geometric decoding. For understanding, the model's text response is evaluated against the observed 3D input under the same critic interface. We compute a scalar rollout reward by aggregating complementary critics:

- **Human Preference Critic** $R_H$: evaluates perceptual preference between the text and multi-view renderings, using renderings from the generated object (generation) or from the input voxel grid (understanding) (Wu et al., 2023; Xu et al., 2023; Ye et al., 2024b).
- **3D Understanding Critic** $R_V$: verifies attribute- and part-level correctness by employing 3D-VQA models (Zhu et al., 2023; Chen et al., 2024b; Qi et al., 2024; Hong et al., 2023). For generation, we instantiate verification queries from $\mathcal{S}_{\text{sem}}^{(i)}$ and score whether the generated shape satisfies the trace; for understanding, we query the 3D-VQA model on rendered views of the input voxel grid and score how well $o_{\text{text}}^{(i)}$ matches the predicted answers.
- **Text-3D Alignment Critic** $R_X$: measures semantic consistency between text and 3D using pretrained text-3D embedding models (Xue et al., 2024; Yin et al., 2025). For generation, $R_X$ aligns $(T_p, \mathcal{S}_{\text{sem}}^{(i)})$ with the generated object $\hat{O}$; for understanding, it aligns $o_{\text{text}}^{(i)}$ with the input 3D observation $V$.
- **Physical Coherence Critic** $R_P$ (generation only): analytic geometry-based reward encouraging (i) stability by penalizing center-of-mass outside the support polygon and insufficient ground contact, (ii) connectivity by penalizing fragmented or weakly attached parts, and (iii) non-interpenetration by penalizing self-intersection and part overlap.

We aggregate normalized critic scores into a rollout reward:

$$R_\tau^{(i)} = \sum_{c \in \mathcal{C}_\tau} w_c R_{c,\tau}^{(i)}, \qquad \mathcal{C}_{\text{gen}} = \{H, V, X, P\}, \qquad \mathcal{C}_{\text{und}} = \{H, V, X\}. \tag{4}$$

**Adaptive Balancing Across Rollout Types.** Because Co-GRPO optimizes two rollout types with different objectives and critic sets, we balance training at the group level using an adaptive mixture. Specifically, we maintain exponential moving averages of the per-mode rollout rewards, $\bar{R}_{\text{gen}}$ and $\bar{R}_{\text{und}}$, and sample the next group type with probability

$$p(\tau) \propto \exp\big(\alpha\,(1 - \bar{R}_\tau)\big), \quad \tau \in \{\text{gen}, \text{und}\}, \tag{5}$$

so the mode with lower recent reward receives more updates. This self-balancing scheme reduces tuning of the generation/understanding ratio and stabilizes optimization under heterogeneous critics.

### 3.4 Training

Training alternates between *generation* and *understanding* updates. At each iteration, we first select a rollout type $\tau \in \{\text{gen}, \text{und}\}$ with probability given by Eq. (5). We then sample a group of $G$ rollouts and perform one Co-GRPO update using within-group advantages. If $\tau = \text{gen}$, given a

Table 1: **3D object captioning results on the Objaverse benchmark.** We evaluate the model's 3D caption capability. CoRe3D achieves state-of-the-art performance by a significant margin across all n-gram and semantic similarity metrics, demonstrating that our reasoning-driven generative training directly enhances 3D understanding. Best and second-best results are highlighted.

| Model | BLEU-1 ↑ | ROUGE-L ↑ | METEOR ↑ | Sentence-BERT ↑ | SimCSE ↑ |
|---|---|---|---|---|---|
| LLaVA-13B | 4.01 | 8.18 | 13.18 | 46.97 | 48.86 |
| Qwen2.5-VL-7B | 4.05 | 7.85 | 14.23 | 48.90 | 50.86 |
| 3D-LLM | 15.11 | 17.84 | 19.22 | 42.36 | 43.58 |
| LEO | 16.98 | 20.12 | 20.91 | 48.01 | 47.25 |
| PointLLM-13B | 3.18 | 7.54 | 12.24 | 47.89 | 49.01 |
| ShapeLLM-Omni | 18.92 | 21.46 | 22.12 | 49.43 | 50.72 |
| **CoRe3D** | 24.02 | 26.45 | 24.98 | 51.17 | 52.79 |

prompt $T_p$ we sample joint rollouts $(\mathcal{S}_{\text{sem}}^{(i)}, \mathcal{G}_{\text{geo}}^{(i)}) \sim \pi_\theta(\cdot \mid T_p)$. Each rollout first produces a semantic trace $\mathcal{S}_{\text{sem}}^{(i)}$, then decodes octant tokens $\mathcal{G}_{\text{geo}}^{(i)}$ with semantic-geometric co-attention, and decodes these tokens into a 3D object for multi-view rendering. If $\tau = \text{und}$, given an observed voxel grid $V$ and query $T_q$, we encode $V$ into octant tokens $\mathcal{G}_{\text{geo}}^{V}$ and sample $G$ semantic traces and responses: $\mathcal{S}_{\text{sem}}^{(i)} \sim \pi_\theta(\cdot \mid T_q, \mathcal{G}_{\text{geo}}^{V})$ and $o_{\text{text}}^{(i)} \sim \pi_\theta(\cdot \mid T_q, \mathcal{G}_{\text{geo}}^{V}, \mathcal{S}_{\text{sem}}^{(i)})$. In both cases, we evaluate the corresponding critic set and aggregate normalized critic scores into rollout rewards $\{R_\tau^{(i)}\}_{i=1}^{G}$. Finally, we compute group-relative advantages from the rewards and update the shared policy parameters $\theta$.

## 4 Experiments

**Implementation Details.** Our training data is sourced from the 3D-Alpaca dataset (Ye et al., 2025b), which comprises approximately 2.56M multimodal samples spanning text-to-3D, image-to-3D, 3D-captioning, and 3D-editing tasks. Each 3D asset is rendered from four orthogonal views and paired with GPT-generated captions, providing rich interleaved supervision across modalities. We initialize our 3D-ULM from ShapeLLM-Omni (Ye et al., 2025b), which we extend with our octant-based 3D VQ-VAE for compact geometry representation and geometric CoT reasoning. Training follows the Co-GRPO framework with a base learning rate of $1 \times 10^{-6}$ and KL regularization $\beta = 0.01$. The model is trained for 40k steps with batch size 256 and group size $G = 8$ on 8×L40 GPUs. We decode predicted octant codes via the 3D VQ-VAE decoder to a voxel grid and extract meshes with Marching Cubes for rendering. For reward computation, we render multi-view images for each generated mesh and evaluate them using HPS (Wu et al., 2023) for the Human Preference Critic, ShapeLLM (Qi et al., 2024) for the 3D Understanding Critic, and ULIP (Xue et al., 2024) embeddings for the Text–3D Alignment Critic. Finally, the Physical Coherence Critic is implemented through geometry-based evaluation on the generated meshes. Additional details can be found in Appendix B.

### 4.1 Quantitative Results

**3D Object Understanding.** We evaluate 3D understanding on the Objaverse (Deitke et al., 2023) held-out set. We compare CoRe3D against general-purpose vision-language models (VLMs) such as LLaVA-13B (Liu et al., 2023a) and Qwen2.5-vl-7B, and specialized 3D-language models 3D-LLM (Hong et al., 2023), LEO (Huang et al., 2024), PointLLM-13B (Xu et al., 2024), and ShapeLLM-Omni (Ye et al., 2025b). Performance is measured using n-gram matching (BLEU-1, ROUGE-L, METEOR) and semantic embedding similarity (Sentence-BERT, SimCSE) following the evaluation settings in PointLLM (Xu et al., 2024). Results in Table 1 show that CoRe3D achieves consistent gains over the strongest baseline on every metric (*e.g.*, BLEU-1 improves ↑ 27.0%, ROUGE-L ↑ 23.3%, and METEOR ↑ 12.9%).

**3D-Generation.** We evaluate the core generative capabilities of CoRe3D on both Text-to-3D and Image-to-3D synthesis. We compare against several leading methods, including SAR3D, CLAY, Trellis, and ShapeLLM-Omni (Chen et al., 2025c; Zhang et al., 2024a; Ye et al., 2025b; Xiang et al., 2024). We use three standard metrics: CLIP Score, Frechet Distance (FD), and Kernel Distance (KD) to assess the generation quality. As shown in Table 2, CoRe3D achieves state-of-the-art Text-to-3D performance, improving over the strongest baseline with CLIP from 0.29 to 0.40 (↑ 37.9%), while also reducing FD and KD. These gains provide quantitative evidence that our Semantic CoT,

Table 2: **Quantitative comparison of 3D generation for Text-to-3D and Image-to-3D tasks.** We evaluate CoRe3D against state-of-the-art generative models, achieving competitive performance across all metrics on both tasks. Best and second-best highlighted.

| Method | Text-to-3D | | | Image-to-3D | | |
|---|---|---|---|---|---|---|
| | CLIP ↑ | $FD_{incep}$ ↓ | $KD_{incep}$ ↓ | CLIP ↑ | $FD_{incep}$ ↓ | $KD_{incep}$ ↓ |
| SAR3D | 0.23 | 28.4 | 0.27 | 0.84 | 22.1 | 0.18 |
| CLAY | 0.27 | 23.9 | 0.21 | 0.85 | 13.5 | 0.10 |
| Trellis | 0.29 | 18.6 | 0.19 | 0.85 | 10.9 | 0.08 |
| ShapeLLM-Omni | 0.27 | 24.4 | 0.24 | 0.84 | 14.1 | 0.09 |
| CoRe3D | 0.40 | 18.5 | 0.18 | 0.86 | 11.2 | 0.08 |

Table 3: **Evaluation of conversational and reasoning abilities on standard language benchmarks.** We compare CoRe3D against general VLMs and 3D-specific language models, with ours achieving competitive results. Best and second-best are highlighted.

| Benchmark | Qwen2.5-vl-7B | LLaMA3.2-Vision-11B | LLaMA-Mesh-8B | ShapeLLM-Omni | CoRe3D |
|---|---|---|---|---|---|
| MMLU ↑ | 67.5 | 66.2 | 59.8 | 64.3 | 67.6 |
| PIQA ↑ | 81.3 | 80.1 | 79.8 | 78.9 | 79.4 |
| GSM8K ↑ | 43.2 | 42.1 | 37.2 | 55.6 | 57.3 |
| SIQA ↑ | 41.0 | 40.6 | 40.3 | 41.5 | 41.5 |

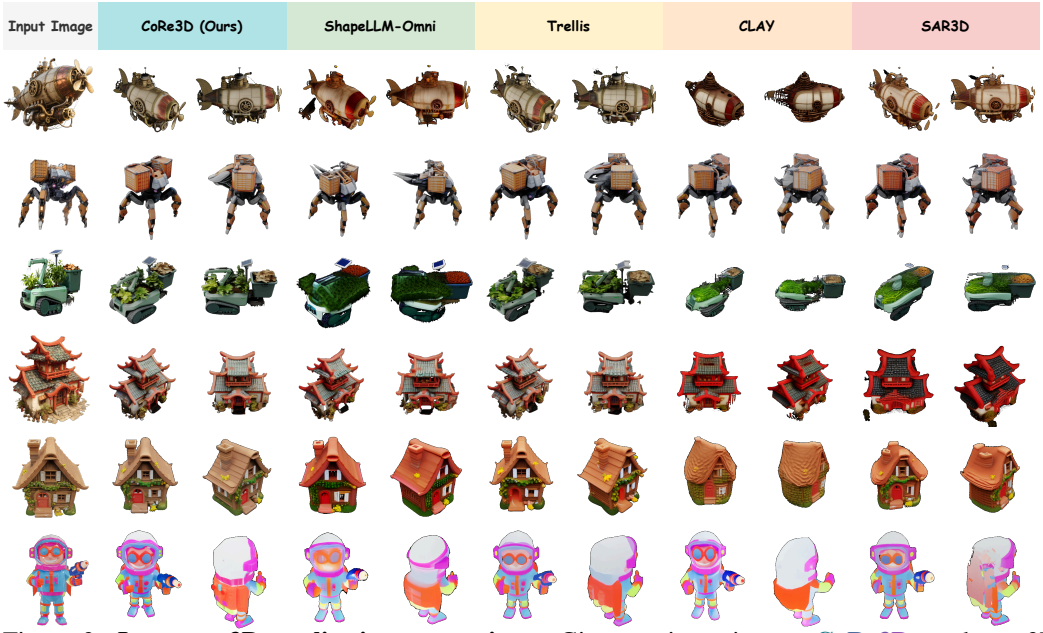



Figure 3: **Image-to-3D qualitative comparison.** Given an input image, CoRe3D produces 3D shapes with higher geometric fidelity, cleaner topology, and stronger semantic alignment.

jointly optimized with Geometric CoT via Co-GRPO, strengthens prompt faithfulness. The improvement extends to Image-to-3D (CLIP at 0.86) while remaining close to the best FD/KD.

**General Conversational Abilities.** We also evaluate the foundational language understanding capabilities of CoRe3D. We compare against leading general-purpose VLMs, such as Qwen2.5-VL-7B (Bai et al., 2025a) and LLaMA3.2-Vision-11B (Touvron et al., 2023), and 3D-focused multimodal models LLaMA-Mesh-8B (Wang et al., 2024b) and ShapeLLM-Omni-7B (Ye et al., 2025b) on a suite of standard language benchmarks. These include MMLU (Hendrycks et al., 2020), PIQA (Bisk et al., 2020), GSM8K (Cobbe et al., 2021), and SIQA (Sap et al., 2019). As shown in Table 3, CoRe3D maintains its general conversational abilities with good language understanding and reasoning performance. These results suggest that our co-reasoning mechanism does not diminish the model's broader reasoning capabilities but instead strengthens them.

### 4.2 Qualitative Results

**Standard 3D Generation.** We first evaluate CoRe3D on standard image-to-3D and text-to-3D tasks, comparing it against strong baselines, including the unified 3D models ShapeLLM-Omni (Ye et al., 2025b) and SAR3D (Chen et al., 2025c), as well as the generation-focused models Trellis (Xiang et al., 2024) and CLAY (Zhang et al., 2024a). As illustrated in Figure 3, competing methods often exhibit misalignment with the input image or suffer from geometric artifacts. In contrast, our model generates 3D meshes with high geometric fidelity and semantic coherence, faithfully capturing the complex structures present in the source image. This superior performance extends to text-to-3D generation, as shown in Figure 4. Our model achieves a more robust alignment with the

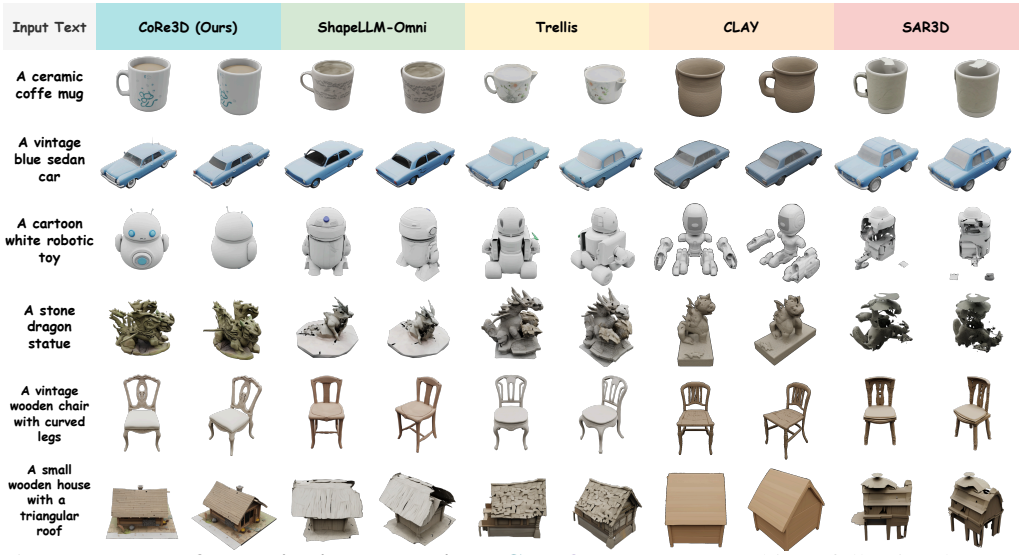



Figure 4: **Text-to-3D qualitative comparison.** CoRe3D generates 3D objects following the prompt.

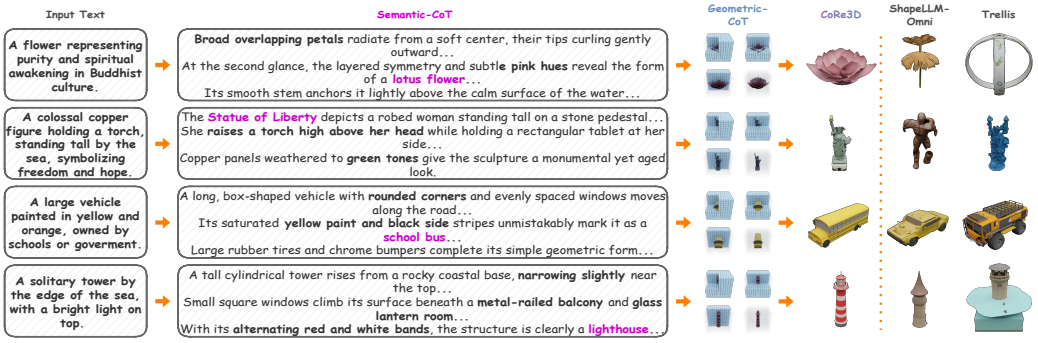



Figure 6: **Qualitative results of CoRe3D on challenging prompts** that require inferring the correct object or interpreting implicit descriptive cues. Comparing with the base model ShapeLLM-Omni (Ye et al., 2025b) and state-of-the-art generation model Trellis (Xiang et al., 2024), our model successfully inferred the true object from the implicit prompts. (*e.g., "a colossal copper figure holding a torch ... symbolizing freedom and hope"* corresponds to *The Statue of Liberty.*)

text prompt. Notably, CoRe3D successfully interprets and renders fine-grained stylistic details, such as the "cartoon" attribute, producing 3D meshes that align precisely with the prompt's intention.

**3D Part Editing.** We additionally evaluate CoRe3D on 3D part editing, where the input is a source object and a natural-language edit request. The edits span diverse categories and operations, including structural changes (e.g., removing armrests, making a backrest taller, replacing legs with wheels) and appearance/attribute edits (e.g., adding a jacket or hat, modifying tail length). As shown in Figure 5, CoRe3D performs targeted modifications while preserving the object's identity and maintaining coherent part geometry. Compared to standard generative models that often introduce global drift or distort unchanged regions, our edits remain localized and produce clean, consistent shapes that reflect the specified change.




Figure 5: **Qualitative results on 3D part editing.** Given a source object and an edit instruction, CoRe3D performs targeted part-level modifications while preserving object identity and maintaining coherent geometry.

Table 4: **Ablation isolating the co-attention interface.** We keep the unified backbone, Co-GRPO, critics, and training schedule fixed, and only change the semantic–geometric coupling.

| Method | 3D Generation | | | 3D Captioning | | | | |
|---|---|---|---|---|---|---|---|---|
| | CLIP↑ | $FD_{incep}$ ↓ | $KD_{incep}$ ↓ | BLEU-1↑ | ROUGE-L↑ | METEOR↑ | SBERT↑ | SimCSE↑ |
| CoRe3D (with co-attn) | 0.40 | 18.5 | 0.18 | 24.02 | 26.45 | 24.98 | 51.17 | 52.79 |
| w/o co-attn | 0.33 | 23.8 | 0.26 | 21.3 | 24.0 | 22.4 | 48.2 | 49.6 |
| Concat fusion (Concat+Proj) | 0.35 | 22.6 | 0.24 | 22.0 | 24.6 | 23.0 | 49.0 | 50.5 |
| Pooled fusion (Pool+Concat+Proj) | 0.36 | 21.7 | 0.23 | 22.8 | 25.3 | 23.8 | 49.9 | 51.4 |

**Reasoning-based 3D Generation.** We evaluate CoRe3D on a challenging set of *reasoning-intensive* text-to-3D prompts that require compositional inference and world knowledge beyond literal keyword matching. Such prompts describe targets indirectly or ambiguously, so the model must infer the intended concept before synthesizing geometry (*e.g.*, inferring "Statue of Liberty" from "a colossal copper figure holding a torch . . . symbolizing freedom and hope"). As shown in Figure 6, standard baselines often respond literally to surface cues, producing semantically mismatched shapes. In contrast, CoRe3D leverages its explicit Semantic CoT to disambiguate intent and guide Geometric CoT decoding, yielding 3D objects that better match the prompt's underlying meaning while maintaining high visual quality and structural coherence.

### 4.3 Ablations

**Co-Attention Ablation.** We keep the unified backbone, Co-GRPO, critics, and training schedule fixed, and only vary the semantic-geometric coupling used in both generation and captioning rollouts. As shown in Table 4, removing co-attention completely degrades both 3D generation (CLIP ↓17.5%, $FD_{incep}$ ↑28.6%) and captioning (BLEU-1 ↓11.3%). Replacing co-attention with attention-free fusion partially recovers performance, with pooled fusion outperforming simple concatenation, but both remain below co-attention.

**Semantic-CoT and Geometric-CoT Components.** We ablate CoRe3D by removing either the Semantic-CoT tokens or the Geometric-CoT tokens for the Co-GRPO pipeline. As shown in Figure 7, removing Semantic-CoT leads to structurally plausible objects but lacks category- and attribute-specific cues. Conversely, removing Geometric-CoT results in objects with clear geometric distortions and simplified shapes.




Figure 7: **Ablating Semantic CoT and Geometric CoT.** Results show both CoT contribute significantly towards the final performance.

**Reward Ablation.** We analyze the role of each critic and their combinations to better understand how different reward signals shape model behavior. As shown in Table 5, each critic contributes complementary improvements. Text–3D alignment yields the largest gains in caption quality, showing its importance for accurate descriptions. 3D Understanding significantly improves generation quality by enforcing stronger object-level structure. Combining multiple critics steadily improves performance: pairs such as (3DU + TA + PC) or (HP + TA + PC) achieve balanced improvements across both captioning and generation tasks.

Table 5: **Ablation of different critics.** HP: Human Preference, 3DU: 3D Understanding, TA: Text-3D Alignment, PC: Physical Coherence. Best and second-best results are highlighted.

| Critic | | | | 3D Captioning | | | 3D Generation | | |
|---|---|---|---|---|---|---|---|---|---|
| HP | 3DU | TA | PC | METEOR ↑ | SBERT ↑ | SimCSE ↑ | CLIP ↑ | $FD_{incep}$ ↓ | $KD_{incep}$ ↓ |
| ✓ | - | - | - | 14.01 | 44.62 | 48.33 | 0.28 | 23.7 | 0.28 |
| - | ✓ | - | - | 14.12 | 47.88 | 49.63 | 0.31 | 22.7 | 0.23 |
| - | - | ✓ | - | 15.34 | 46.52 | 50.11 | 0.33 | 21.9 | 0.21 |
| - | - | - | ✓ | 13.41 | 45.76 | 48.62 | 0.30 | 20.8 | 0.22 |
| - | ✓ | ✓ | ✓ | 18.92 | 48.77 | 51.24 | 0.36 | 19.7 | 0.19 |
| ✓ | - | ✓ | ✓ | 20.48 | 49.02 | 51.89 | 0.37 | 19.9 | 0.20 |
| ✓ | ✓ | - | ✓ | 19.83 | 49.88 | 51.01 | 0.38 | 20.1 | 0.19 |
| ✓ | ✓ | ✓ | - | 21.55 | 50.41 | 51.43 | 0.37 | 19.4 | 0.20 |
| ✓ | ✓ | ✓ | ✓ | 24.98 | 51.17 | 52.79 | 0.40 | 18.5 | 0.18 |

## 5 Conclusion

We introduce CoRe3D, a collaborative reasoning framework that unifies semantic and geometric reasoning through a dual chain-of-thought process. Beyond producing coherent 3D assets, CoRe3D demonstrates robust interpretive capabilities, handling indirect and referring descriptions. Our results highlight collaborative reasoning as a scalable foundation for general 3D intelligence.

### AI USE STATEMENT

In this work, we used generative AI tools mainly to assist in editing to improve readability. We have not used generative AI tools in the development of the research contributions, method implementation, result interpretation, or literature search.

### ETHICS STATEMENT

CoRe3D advances unified 3D intelligence by enabling models to both *interpret* and *construct* 3D objects through collaborative semantic and geometric reasoning. Such capabilities can benefit a wide spectrum of applications. In robotics and embodied AI, improved spatial reasoning may support better manipulation, affordance understanding, and task planning. In simulation and digital twin systems, controllable 3D generation can streamline asset creation and accelerate workflows in engineering and education. At the same time, such frameworks carry several potential risks. Unified 3D models may unintentionally reproduce private or proprietary content, and generative pipelines could be misused to create realistic yet deceptive 3D assets. CoRe3D is intended for research and educational use, and we encourage responsible deployment that emphasizes transparency, provenance tracking, consent for training data, and adherence to domain-specific usage guidelines. With such measures in place, we believe the benefits of collaborative reasoning for 3D understanding and generation outweigh the potential risks.

### REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed implementation specifications, including hyperparameter settings and evaluation protocols. We will also publicly release model checkpoints, training, and evaluation code.

## REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *2502.13923*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

David Benson and Joel Davis. Octree textures. *ACM Transactions on Graphics (TOG)*, 21(3): 785–790, 2002.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, volume 34, pp. 7432–7439, 2020.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. `https://github.com/Deep-Agent/R1-V`, 2025a. Accessed: 2025-02-02.

Luxi Chen, Zhengyi Wang, Zihan Zhou, Tingting Gao, Hang Su, Jun Zhu, and Chongxuan Li. Microdreamer: Efficient 3d generation in $\sim$20 seconds by score-based iterative reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, pp. 22246–22256, 2023.

Sijin Chen, Xin Chen, Anqi Pang, Xianfang Zeng, Wei Cheng, Yijun Fu, Fukun Yin, Zhibin Wang, Jingyi Yu, Gang Yu, et al. Meshxl: Neural coordinate field for generative 3d foundation models. *NeurIPS*, 37:97141–97166, 2024a.

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*, pp. 26428–26438, 2024b.

Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, et al. Meshanything: Artist-created mesh generation with autoregressive transformers. *arXiv preprint arXiv:2406.10163*, 2024c.

Yiwen Chen, Yikai Wang, Yihao Luo, Zhengyi Wang, Zilong Chen, Jun Zhu, Chi Zhang, and Guosheng Lin. Meshanything v2: Artist-created mesh generation with adjacent mesh tokenization. *arXiv preprint arXiv:2408.02555*, 2024d.

Yongwei Chen, Yushi Lan, Shangchen Zhou, Tengfei Wang, and Xingang Pan. Sar3d: Autoregressive 3d object generation and understanding via multi-scale 3d vqvae. In *CVPR*, 2025c.

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. *arXiv preprint arXiv:2409.12957*, 2024e.

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*, pp. 24185–24198, 2024f.

Zilong Chen, Feng Wang, Yikai Wang, and Huaping Liu. Text-to-3d using gaussian splatting. In *CVPR*, pp. 21401–21412, 2024g.

Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024h.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *NeurIPS*, 36:35799–35813, 2023.

Kangle Deng, Hsueh-Ti Derek Liu, Yiheng Zhu, Xiaoxia Sun, Chong Shang, Kiran S Bhat, Deva Ramanan, Jun-Yan Zhu, Maneesh Agrawala, and Tinghui Zhou. Efficient autoregressive shape generation via octree-based adaptive tokenization. In *ICCV*, pp. 11685–11696, 2025.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

Zhirui Gao, Renjiao Yi, Yuhang Huang, Wei Chen, Chenyang Zhu, and Kai Xu. Partgs: Learning part-aware 3d representations by fusing 2d gaussians and superquadrics. *arXiv preprint arXiv:2408.10789*, 2024.

Ethan Griffiths, Maryam Haghighat, Simon Denman, Clinton Fookes, and Milad Ramezani. Hotformerloc: Hierarchical octree transformer for versatile lidar place recognition across ground and aerial views. In *CVPR*, pp. 6648–6658, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zekun Hao, David W Romero, Tsung-Yi Lin, and Ming-Yu Liu. Meshtron: High-fidelity, artist-like 3d mesh generation at scale. *arXiv preprint arXiv:2412.09548*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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*, 36:20482–20494, 2023.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world, 2024.

Ting Huang, Zeyu Zhang, and Hao Tang. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. *arXiv preprint arXiv:2507.23478*, 2025a.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025b.

Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M Rehg, and Varun Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. *arXiv preprint arXiv:2501.04689*, 2025c.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arxiv:2310.02596*, 2023.

Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024.

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*, pp. 300–309, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model, 2024a.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023a.

Jiawei Liu, Nirav Diwan, Zhe Wang, Haoyu Zhai, Xiaona Zhou, Kiet A Nguyen, Tianjiao Yu, Muntasir Wahed, Yinlin Deng, Hadjer Benkraouda, et al. Purpcode: Reasoning for safer code generation. *arXiv preprint arXiv:2507.19060*, 2025.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36:44860–44879, 2023b.

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. *NeurIPS*, 36:22226–22246, 2023c.

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*, pp. 9298–9309, 2023d.

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. *arXiv preprint arXiv:2309.03453*, 2023e.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

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*, pp. 9970–9980, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. 2024:23439–23554, 2024.

Sining Lu, Guan Chen, Nam Anh Dinh, Itai Lang, Ari Holtzman, and Rana Hanocka. Ll3m: Large language 3d modelers. *arXiv preprint arXiv:2508.08228*, 2025.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815–824, 2023.

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

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

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 *CVPR*, pp. 9914–9925, 2024.

Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. In *CVPR*, pp. 2349–2359, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, pp. 3577–3586, 2017.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

Ruwen Schnabel and Reinhard Klein. Octree-based point-cloud compression. *PBG@ SIGGRAPH*, 2006.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *arXiv preprint arXiv:2506.21656*, 2025.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023a.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023b.

Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *CVPR*, pp. 19615–19625, 2024.

Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023.

Richard Szeliski. Rapid octree construction from image sequences. *CVGIP: Image understanding*, pp. 23–32, 1993.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

Jiaxiang Tang, Zhaoshuo Li, Zekun Hao, Xian Liu, Gang Zeng, Ming-Yu Liu, and Qinsheng Zhang. Edgerunner: Auto-regressive auto-encoder for artistic mesh generation. *arXiv preprint arXiv:2409.18114*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. doi: 10.48550/arXiv.2405.09818.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2024.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022.

Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.

Peng-Shuai Wang. Octformer: Octree-based transformers for 3d point clouds. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023.

Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions On Graphics (TOG)*, pp. 1–11, 2017.

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, pp. 4563–4573, 2023a.

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*, 2024a.

Xinzhou Wang, Yikai Wang, Junliang Ye, Zhengyi Wang, Fuchun Sun, Pengkun Liu, Ling Wang, Kai Sun, Xintong Wang, and Bin He. Animatabledreamer: Text-guided non-rigid 3d model generation and reconstruction with canonical score distillation. *arXiv preprint arXiv:2312.03795*, 2023b.

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*, 36:8406–8441, 2023c.

Zhengyi Wang, Jonathan Lorraine, Yikai Wang, Hang Su, Jun Zhu, Sanja Fidler, and Xiaohui Zeng. Llama-mesh: Unifying 3d mesh generation with language models. *arXiv preprint arXiv:2411.09595*, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837, 2022.

Si-Tong Wei, Rui-Huan Wang, Chuan-Zhi Zhou, Baoquan Chen, and Peng-Shuai Wang. Octgpt: Octree-based multiscale autoregressive models for 3d shape generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

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.

Haohan Weng, Yikai Wang, Tong Zhang, CL Chen, and Jun Zhu. Pivotmesh: Generic 3d mesh generation via pivot vertices guidance. *arXiv preprint arXiv:2405.16890*, 2024a.

Haohan Weng, Zibo Zhao, Biwen Lei, Xianghui Yang, Jian Liu, Zeqiang Lai, Zhuo Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, et al. Scaling mesh generation via compressive tokenization. *arXiv preprint arXiv:2411.07025*, 2024b.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.

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. *arXiv preprint arXiv:2405.14832*, 2024b.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *CVPR*, pp. 4840–4851, 2024c.

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. *arXiv preprint arXiv:2412.01506*, 2024.

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*, 2024.

Bojun Xiong, Jialun Liu, Jiakui Hu, Chenming Wu, Jinbo Wu, Xing Liu, Chen Zhao, Errui Ding, and Zhouhui Lian. Texgaussian: Generating high-quality pbr material via octree-based 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 551–561, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS*, 36:15903–15935, 2023.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pp. 131–147. Springer, 2024.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *CVPR*, pp. 27091–27101, 2024.

Yutaro Yamada, Khyathi Chandu, Bill Yuchen Lin, Jack Hessel, Ilker Yildirim, and Yejin Choi. L3go: Language agents with chain-of-3d-thoughts for generating unconventional objects. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 456–469, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024b.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 2024a.

Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. In *ICCV*, 2025a.

Junliang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun Zhu. Dreamreward: Text-to-3d generation with human preference. In *European Conference on Computer Vision*, pp. 259–276. Springer, 2024b.

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*, 2025b.

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*, pp. 6796–6807, 2024.

Fukun Yin, Xin Chen, Chi Zhang, Biao Jiang, Zibo Zhao, Wen Liu, Gang Yu, and Tao Chen. Shapegpt: 3d shape generation with a unified multi-modal language model. *IEEE Transactions on Multimedia*, 2025.

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *ICCV*, pp. 5752–5761, 2021.

Tianjiao Yu, Vedant Shah, Muntasir Wahed, Ying Shen, Kiet A Nguyen, and Ismini Lourentzou. Part$^2$GS: Part-aware modeling of articulated objects using 3d gaussian splatting. *arXiv preprint arXiv:2506.17212*, 2025.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

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, 2024a.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024b.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024c.

Ruowen Zhao, Zhengyi Wang, Yikai Wang, Zihan Zhou, and Jun Zhu. Flexidreamer: single image-to-3d generation with flexicubes. *arXiv preprint arXiv:2404.00987*, 2024.

Ruowen Zhao, Junliang Ye, Zhengyi Wang, Guangce Liu, Yiwen Chen, Yikai Wang, and Jun Zhu. Deepmesh: Auto-regressive artist-mesh creation with reinforcement learning. *arXiv preprint arXiv:2503.15265*, 2025.

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. *NeurIPS*, 36:73969–73982, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *ICCV*, pp. 2911–2921, 2023.

# APPENDIX

## A ADDITIONAL RELATED WORK

**Post-Pretraining Reinforcement Learning.** Since the introduction of Chain-of-Thought prompting (Wei et al., 2022), enhancing the reasoning ability of large language models has become an important research focus. More recently, DeepSeek-R1 (Guo et al., 2025) advanced this direction by proposing a rule-based reward design combined with GRPO training (Shao et al., 2024), encouraging models to produce explicit intermediate reasoning traces before generating final answers. This paradigm has since been extended to multimodal settings (Meng et al., 2025; Yang et al., 2025; Zhang et al., 2025; Huang et al., 2025b), where task-specific rewards guide the learning process. These reasoning-driven methods have achieved notable gains across a variety of challenging tasks (Liu et al., 2024c), including mathematical problem solving (Zhang et al., 2024b;c) and code generation (Jain et al., 2024; Liu et al., 2025). Unlike prior GRPO-based methods that optimize textual or multimodal response trajectories separately, our proposed **CoRe3D** employs a collaborative reasoning strategy that tightly couples linguistic semantics with 3D geometry, leading to a coherent 3D unified model for both 3D generation and 3D understanding.

**Octant-based 3D Representations.** The concept of octree has been used in a wide range of 3D geometric processing applications, including point cloud compression (Schnabel & Klein, 2006), 3D texturing (Benson & Davis, 2002; Xiong et al., 2025), and multi-view scene reconstruction (Szeliski, 1993; Yu et al., 2021). Beyond these foundational uses, octree structures have become integral to efficient shape analysis in large-scale environments (Wang et al., 2017; Riegler et al., 2017; Wang, 2023; Wu et al., 2024c; Griffiths et al., 2025). In the realm of generative modeling, recent advancements (Deng et al., 2025; Wei et al., 2025) employ adaptive tokenization or multi-scale autoregressive strategies to allocate computational resources to geometrically complex regions dynamically. Unlike prior methods, we introduce an octant-based 3D VQ-VAE that discretizes the 3D volume into uniform blocks, providing a stable substrate for geometric reasoning. This design preserves essential spatial locality while maintaining a compact token sequence, thereby facilitating stable and interpretable geometric chain-of-thought reasoning within a unified 3D-LLM framework.

## B METHOD DETAILS

**Octant-based Autoregressive Model.** As shown in Figure 8, our voxel VQ-VAE maps each $64^3$ input volume to a $16^3$ latent grid. To reduce sequence length while preserving local geometric structure, we partition this latent grid into non-overlapping $2\times2\times2$ neighborhoods. Each neighborhood forms a local *octant block* that contains eight spatially adjacent latent cells. Concatenating the features of these eight cells produces a single octant token, yielding exactly $8\times8\times8 = 512$ tokens per object. These tokens summarize compact spatial regions and maintain local geometry and appearance cues.

To serialize the latent volume into an autoregressive sequence, we employ a Morton (Z-order) space-filling curve, which preserves spatial locality more effectively than raster or lexicographic scanning. Let $\mathcal{G}_{geo} = [g_1, g_2, \ldots, g_{512}]$ denote the sequence of octant tokens arranged in Morton order. The generative process factorizes the distribution over the latent volume as

$$p(\mathcal{G}_{geo}) = \prod_{i=1}^{512} p(g_i \mid g_{<i}, T_p, \mathcal{S}_{\text{sem}}), \tag{6}$$

where the semantic chain-of-thought $\mathcal{S}_{\text{sem}}$ provides high-level cues that guide geometric synthesis. Each factor predicts the code for the next spatial block using the prompt, the semantic trace, and the preceding geometric codes. After the complete sequence has been generated, the VQ-VAE decoder reconstructs the voxel field. The geometric trace therefore records the sequence of spatial code predictions used to construct the object.

To encode spatial location, we attach a learned positional embedding to each octant token, keyed by its block index $(x_b, y_b, z_b)$ in the $8\times8\times8$ grid. This embedding is injected *after* vector quantization, ensuring that the codebook remains content-centric while the autoregressive decoder remains location-aware. Together, the octant blocks and positional embeddings introduce two inductive bi-

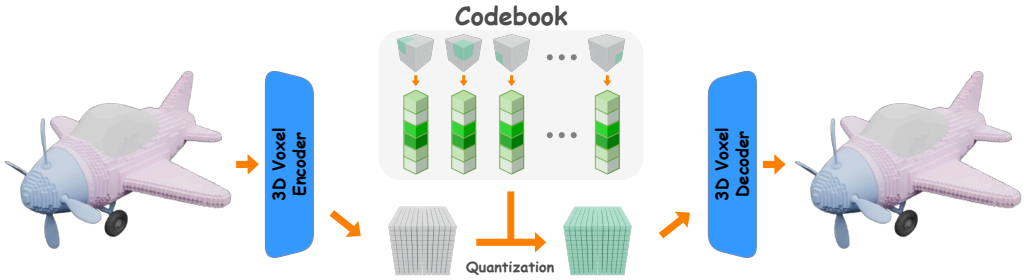



Figure 8: **Octant-based 3D VQVAE.** The voxelized geometry is encoded into latent blocks, quantized using a shared codebook of 3D embeddings, and decoded back into a high-fidelity voxel grid.

ases: (i) each token carries high-resolution local geometric context, and (ii) the sequence ordering respects 3D locality.

During training, the semantic reasoning trace $\mathcal{S}_{\text{sem}}$ and the geometric trace $\mathcal{G}_{\text{geo}}$ jointly condition the autoregressive transformer. The semantic trace encodes global planning information such as categories and textures, while the geometric trace provides localized structural cues for neighborhood-level voxel refinement. The decoder predicts a discrete codebook index for each octant token, reconstructing the latent volume block-by-block. After predicting all 512 tokens, the VQ-VAE decoder reconstructs a dense $64^3$ voxel field, which is subsequently rendered into mesh or multi-view images. Compared to hierarchical octree models (Wei et al., 2025), which must autoregressively generate thousands of binary split or leaf tokens, our design yields a significantly shorter and more expressive sequence. This compact, locality-aware autoregressive formulation is essential to our framework, enabling efficient token-level generation that tightly aligns with semantic and geometric reasoning.

**Physical Coherence Critic.** Whereas the other critics directly use pretrained evaluators, the physical coherence critic is computed from explicit geometric criteria. We use Trimesh to compute physical statistics and PyMeshFix to diagnose self-intersections. Stability is estimated by projecting the mesh center of mass onto the ground plane and checking whether this projection lies inside the convex hull of bottom support vertices extracted from the lowest-$z$ region, yielding a normalized stability score. Structural connectivity is measured by splitting the mesh into connected components and computing the fraction of faces belonging to the largest component, which penalizes fragmented or floating parts. To assess self-intersection, we run PyMeshFix on the original mesh and compare the number of faces before and after repair; a larger relative change indicates more severe geometric artifacts and results in a lower self-intersection score.

## C Implementation Detils

**Hyperparameters.** The policy is updated using KL-regularized GRPO with ratio clipping. The KL penalty coefficient is set to $\beta = 0.01$ and the clipping threshold to $\varepsilon$=0.1. For each prompt, we sample $K$=4 candidate generations and compute pairwise preferences using the four critics introduced in Section 3. The human-preference, 3D-understanding, Text-3D-alignment, and Physical coherence critics are weighted by $(w_H, w_V, w_X, w_P) = (0.25, 0.25, 0.25, 0.25)$. Optimization uses AdamW with a learning rate of $1 \times 10^{-6}$, $\beta_1$=0.9, $\beta_2$=0.98, and weight decay 0.01. We apply a 2000-step linear warmup followed by cosine decay. For stability, the policy KL is capped at $1.2\times$ its exponential moving average, and gradients are clipped with a global norm of 1.0.

**Evaluation Metrics.** We report common text-generation metrics, *i.e.*,, BLEU-1 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee & Lavie, 2005). Although standard, these metrics often favor shorter outputs and may overlook semantic fidelity. To address this, we incorporate embedding-based similarity measures, including Sentence-BERT (Reimers & Gurevych, 2019) and SimCSE (Gao et al., 2021), which evaluate semantic alignment between generated captions and human references more robustly.

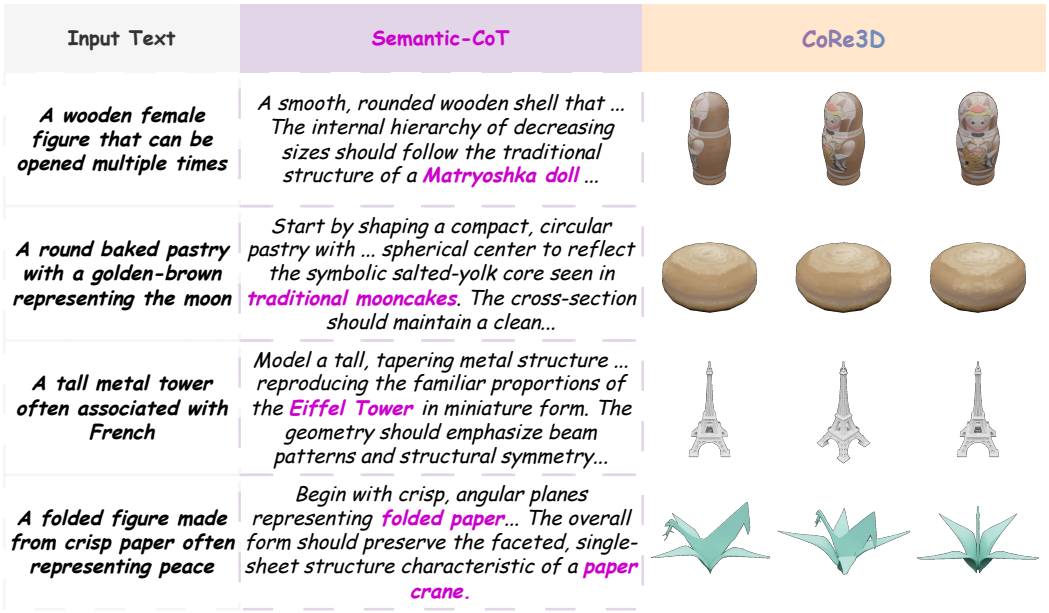



Figure 9: **Additional Results on Challenging Prompts.** CoRe3D successfully inferred the true object from the implicit prompts.

Table 6: **Ablation of Semantic and Geometric CoT.** Best and second-best highlighted.

| | Critic | | 3D Captioning | | | 3D Generation | | |
|---|---|---|---|---|---|---|---|---|
| **Model Name** | **Semantic CoT** | **Geometric CoT** | METEOR ↑ | Sentence-BERT ↑ | SimCSE ↑ | CLIP ↑ | $FD_{incep}$ ↓ | $KD_{incep}$ ↓ |
| ShapeLLM-Omni | ✗ | ✗ | 12.76 | 44.91 | 48.14 | 0.29 | 23.4 | 0.25 |
| CoRe3D | ✓ | ✗ | 16.42 | 50.38 | 51.14 | 0.32 | 21.6 | 0.21 |
| CoRe3D | ✗ | ✓ | 14.89 | 47.32 | 50.41 | 0.35 | 19.8 | 0.19 |
| CoRe3D | ✓ | ✓ | 17.03 | 49.67 | 52.11 | 0.38 | 18.7 | 0.17 |

## D Additional Experiments

**Additional Results on Challenging Prompts.** Figure 9 presents additional qualitative examples demonstrating CoRe3D's ability to infer the correct 3D object even when the input prompt provides only indirect or symbolic descriptions. These prompts intentionally avoid naming the target object and instead describe cultural context, functional cues, or high-level visual impressions. Despite this ambiguity, CoRe3D consistently recovers the correct underlying structure. As shown, the model not only identifies the implicit object but also reconstructs a spatially coherent and visually faithful 3D shape. To our knowledge, CoRe3D is the first 3D generative framework capable of resolving such implicit, referential prompts through semantic reasoning, rather than relying on explicit object mentions or category supervision.

**of Semantic-CoT and Geometric CoT Ablation.** We conduct a quantitative analysis to isolate the contributions of the two complementary reasoning CoTs. As shown in Table 6, introducing **Semantic CoT** yields the largest gains in 3D captioning quality, improving METEOR, Sentence-BERT, and SimCSE scores by a significant margin over ShapeLLM-Omni. This indicates that explicit semantic reasoning helps the model better decompose text prompts into linguistically grounded object attributes. In contrast, **Geometric CoT** primarily benefits 3D generation. By encouraging structured spatial reasoning over octant-level geometry, it substantially improves CLIP similarity and reduces both FD and KD, demonstrating its effectiveness in guiding the model toward globally coherent 3D shapes. Our full model, CoRe3D, achieves the best overall performance across all metrics. These results verify that collaborative reasoning is essential for robust 3D understanding and generation.

**Optimal Codebook Size.** Table 7 evaluates how the capacity of the VQ-VAE codebook affects both semantic alignment and 3D generation fidelity. A small codebook (2048 entries) creates a severe quantization bottleneck. This under-expressiveness leads to noticeably lower METEOR, SBERT, and SimCSE scores, along with substantially worse FD/KD metrics. Increasing the codebook to

Table 7: **Ablation of Codebook Size in the Octant-based 3D VQ-VAE.** Larger codebooks reduce quantization error but exhibit diminishing returns and instability at extreme scales. Octant depth is fixed at 512 octants (Depth 3). Best and second-best.

| Architecture | | 3D Captioning | | | 3D Generation | | |
|---|---|---|---|---|---|---|---|
| **Codebook Size** | **Octants** | METEOR ↑ | SBERT ↑ | SimCSE ↑ | CLIP ↑ | $FD_{incep}$ ↓ | $KD_{incep}$ ↓ |
| **2048** | 512 | 13.81 | 46.92 | 47.44 | 0.31 | 27.8 | 0.24 |
| **4096** | 512 | 15.92 | 48.51 | 50.03 | 0.34 | 22.1 | 0.20 |
| **8192** | 512 | 17.03 | 49.67 | 52.11 | 0.38 | 18.7 | 0.17 |
| **16384** | 512 | 16.77 | 49.98 | 51.62 | 0.37 | 19.3 | 0.18 |

Table 8: **Ablation of Octant Depth in the Octant-based 3D VQ-VAE.** Increasing depth increases spatial granularity ($8 \rightarrow 4096$ octants) and improves fidelity up to Depth 3, after which we observe diminishing returns. Best and second-best.

| Architecture | | 3D Captioning | | | 3D Generation | | |
|---|---|---|---|---|---|---|---|
| **Depth** | **# Octants** | METEOR ↑ | SBERT ↑ | SimCSE ↑ | CLIP ↑ | $FD_{incep}$ ↓ | $KD_{incep}$ ↓ |
| **1** | 8 | 7.34 | 41.82 | 43.10 | 0.21 | 42.3 | 0.34 |
| **2** | 64 | 12.87 | 46.51 | 47.88 | 0.29 | 29.4 | 0.25 |
| **3** | 512 | 17.03 | 49.67 | 52.11 | 0.38 | 18.7 | 0.17 |
| **4** | 4096 | 16.44 | 49.92 | 51.36 | 0.37 | 19.0 | 0.18 |

4096 entries alleviates this issue and yields consistent gains across all evaluation dimensions. Our default setting with an 8192-entry codebook achieves the best overall performance. Expanding the codebook further to 16384 entries results in only marginal improvements on SBERT and CLIP similarity, suggesting limited additional benefit beyond 8192 entries. These findings indicate that our 8192-entry codebook is optimal.

**Optimal Octant Layers.** Table 8 evaluates how the choice of octant depth and therefore the total number of octant tokens affects the performance of our octant-based 3D VQ-VAE. Increasing the depth refines the spatial partitioning of the latent volume (from 8 to 4096 octants), granting the model access to progressively finer geometric detail. We observe a consistent upward trend from Depth 1 to Depth 3 across all captioning and generation metrics. Depth 1 (only 8 octants) severely under-parameterizes local geometry, leading to weak semantic alignment and the worst FD/KD scores. Depth 2 (64 octants) recovers coarse global structure but still lacks local detail, resulting in moderate improvements. Depth 3 provides the best balance between spatial granularity and autoregressive stability. It achieves the highest scores for METEOR, SimCSE, CLIP, and the lowest FD/KD, demonstrating that this level of decomposition offers enough local resolution to capture high-frequency geometry without excessively lengthening the autoregressive sequence. We use Depth 3 as our default. At Depth 4, the finer subdivision (4096 octants) brings only marginal gains in SBERT similarity and CLIP alignment, suggesting diminishing returns beyond Depth 3.

**Additional Image-to-3D Results.** Figure 10 shows additional qualitative results from our image-to-3D pipeline. Despite being a unified model rather than a specialized reconstruction system, CoRe3D generates stable and coherent 3D shapes even from visually complex or cluttered inputs. The model maintains globally consistent geometry, plausible spatial structure, and strong color fidelity, delivering reconstructions that capture the essential form, material cues, and overall visual identity of the input objects.

**Comparison with Zero-shot CoT.** Figure 11 compares our trained semantic-level CoT with a zero-shot CoT baseline, Qwen2.5-VL-7B (Bai et al., 2025b). In the zero-shot setting, the model is prompted to produce a free-form structural description before generating the 3D object, but this unguided semantic reasoning provides only shallow, low-information descriptions. For instance, in the "compact car" example, the zero-shot CoT yields a generic outline of a smooth white prototype with no color, style cues, or structural modifiers, leading to an oversimplified reconstruction. In contrast, the CoRe3D semantic-level CoT produces richer and more actionable structural cues, such as bright blue body color, semi-transparent windows, stylized panel lines, or the presence of a roof rack with parallel bars and a fin-like rear element (bottom CoT row), and the model correspondingly reconstructs a much more faithful 3D shape.

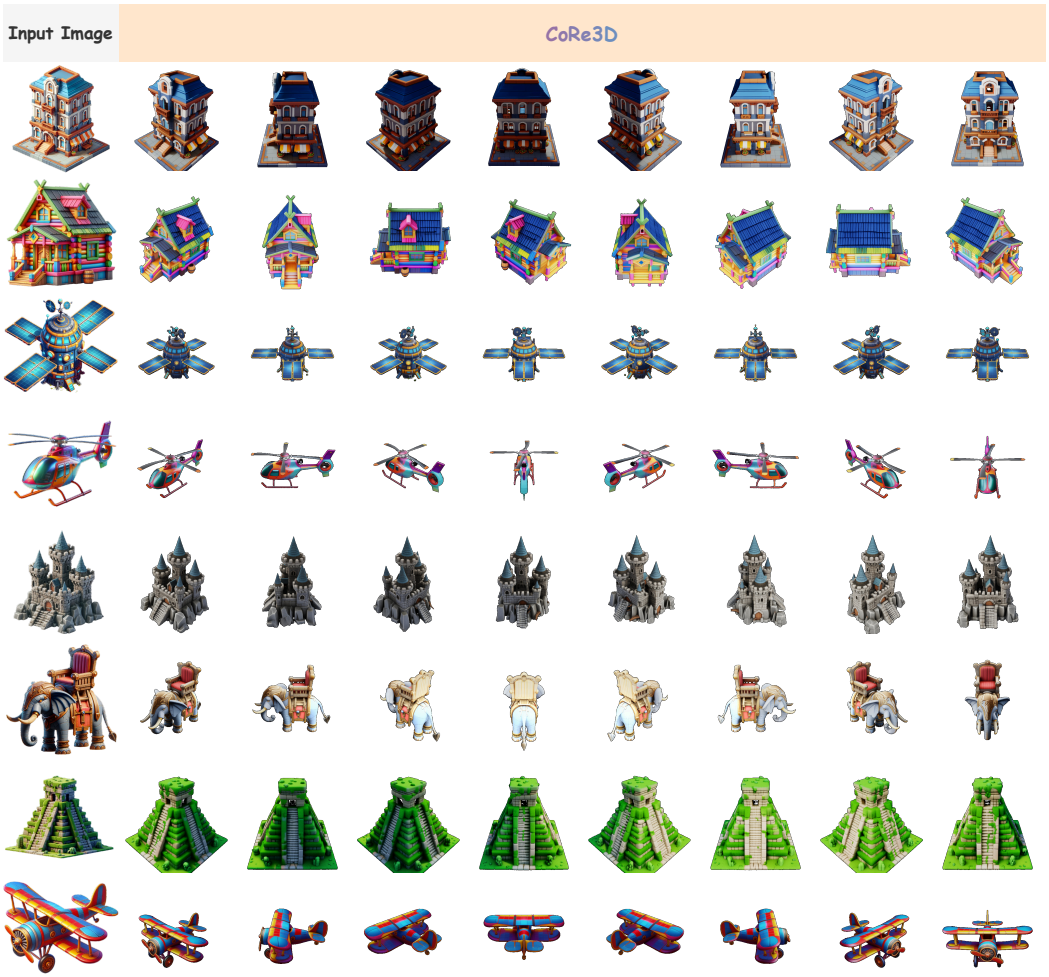



Figure 10: **Additional Image-to-3D Results.** We use visually complex image prompts to demonstrate that CoRe3D achieves strong 3D reconstruction capability, producing coherent geometry and spatially consistent shapes even for challenging inputs.

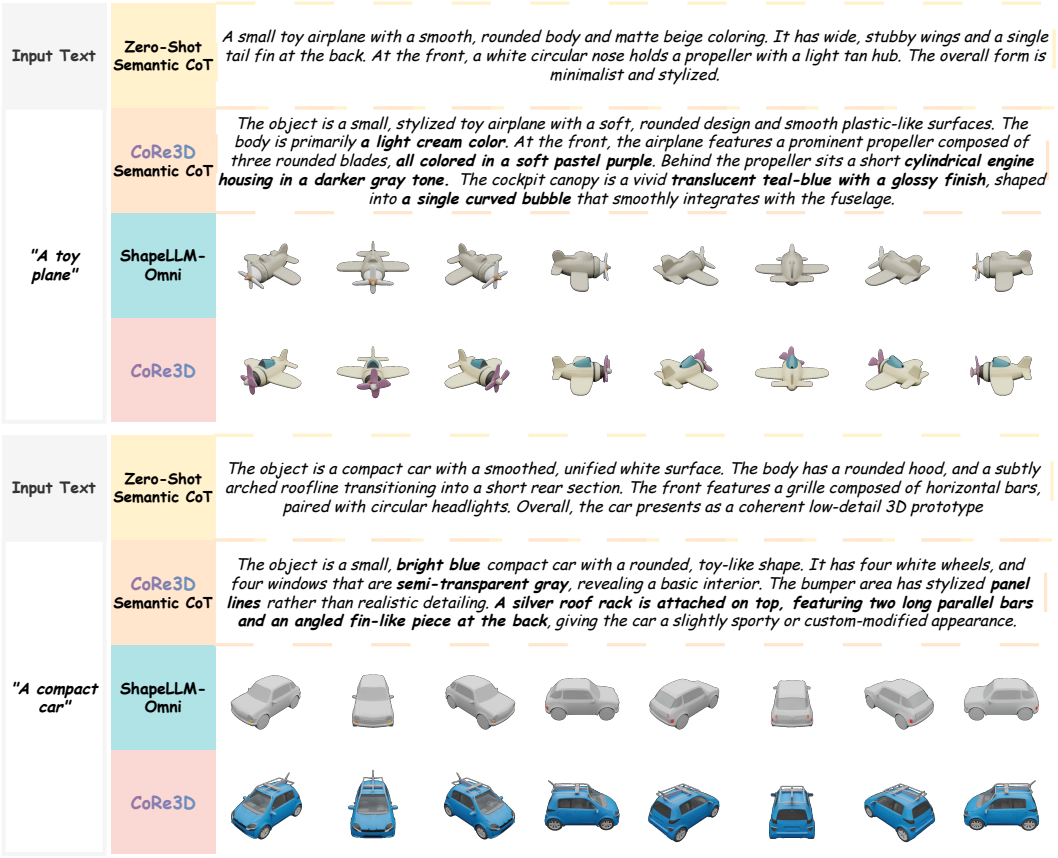



Figure 11: **Comparison with Zero-shot CoT.** We compare our trained semantic-level CoT against a zero-shot CoT baseline (Qwen2.5-VL-7B), where the model is prompted to write a free-form description before generation. Zero-shot CoT tends to be shallow and generic, which yields over-simplified reconstructions. In contrast, CoRe3D produces richer, more actionable semantic cues that translate into more faithful 3D shapes.