DreamPartGen: Semantically Grounded Part-Level 3D Generation via Collaborative Latent Denoising

Tianjiao Yu[1], Xinzhuo Li[1], Muntasir Wahed[1], Jerry Xiong[1], Yifan Shen[1], Ying Shen[1], Ismini Lourentzou[1]

University of Illinois Urbana-Champaign
{ty41, lourent2}@illinois.edu

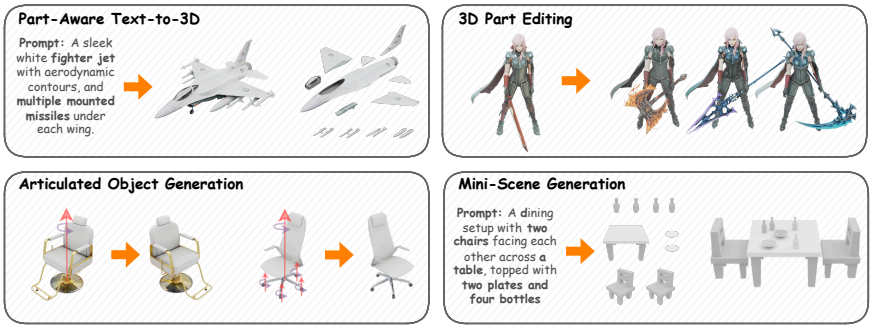

Fig. 1: DreamPartGen connects part-level geometry and appearance with language-driven relational semantics, providing precise control over how parts are modified, arranged, and contextualized. This unified representation enables a wide range of text-to-3D applications, including fine-grained part editing, articulated object generation, and mini-scene synthesis.

Abstract. Understanding and generating 3D objects as compositions of meaningful parts is fundamental to human perception and reasoning. However, most text-to-3D methods overlook the semantic and functional structure of parts. While recent part-aware approaches introduce decomposition, they remain largely geometry-focused, lacking semantic grounding and failing to model how parts align with textual descriptions or their inter-part relations. We propose DreamPartGen, a framework for semantically grounded, part-aware text-to-3D generation. DreamPartGen introduces Duplex Part Latents (DPLs) that jointly model each part's geometry and appearance, and Relational Semantic Latents (RSLs) that capture inter-part dependencies derived from language. A synchronized co-denoising process enforces mutual geometric and semantic consistency, enabling coherent, interpretable, and text-aligned 3D synthesis. Across multiple benchmarks, DreamPartGen delivers state-of-the-art performance in geometric fidelity (↓60% Chamfer Distance) and text–shape alignment (≥↑20% CLIP/ULIP), while producing compositionally consistent and controllable parts.
PLAN Lab https://plan-lab.github.io/dreampartgen

Keywords: Part-Aware 3D Generation · Language-Grounded 3D Generation · Collaborative Part-Latent Denoising

1 Introduction

Many text prompts for 3D generation specify not only *what* parts an object has, but *how* they relate (*e.g.*, a handle *attached to* a mug, wheels *symmetric* on a chassis, a lid *on top of* a box). Capturing these part-level relations is crucial for controllable generation and downstream use cases such as part editing and articulated synthesis [19, 32, 53]. However, most text-to-3D methods operate on *monolithic* latents that entangle geometry, appearance, and semantics, with no explicit representation of part identities or inter-part relations [6, 23, 24, 28, 35, 40, 45]. Recent part-aware methods take a step forward by synthesizing objects from part primitives guided by part segmentations or bounding boxes [4, 12, 18, 26, 49]. Although these approaches improve geometric granularity, they are still brittle to segmentation noise and can be difficult to scale across diverse categories and prompts. More importantly, many part-based frameworks still treat parts as geometrically isolated units. They do not model inter-part relations as explicit variables, and language remains largely non-operational. As a result, part-aware generation is not yet *language-grounded assembly*.

We argue that true part-aware text-to-3D generation requires a different abstraction: a semantically grounded representation in which parts are meaningful entities, and language provides relational structure in addition to describing appearance. Concretely, we introduce **DreamPartGen**, a language-grounded, collaborative part-latent diffusion framework that treats compositional semantics as an explicit representation during denoising. Each object is encoded into **Duplex Part Latents (DPLs)**, paired 3D and 2D latent sequences that jointly capture a part's geometry and appearance, whereas a learnable identifier embedding preserves slot identity across timesteps and instances, keeping parts trackable throughout diffusion. In parallel, we introduce **Relational Semantic Latents (RSLs)**, compact text-derived latents that encode part-level attributes and inter-part relations. Rather than using language only as one-shot conditioning, DreamPartGen performs synchronized co-denoising: DPLs and RSLs co-evolve through part-level and object-level synchronization so that geometry and appearance are refined under persistent, language-derived relational guidance, enforcing mutual geometric–semantic consistency.

To enable supervision at scale, we curate **PartRel3D**, a large-scale relational dataset that augments each object with canonicalized functional and spatial triplets linking parts through explicit semantic predicates. These canonicalized relations are encoded into RSLs, allowing the model to learn assembly-level consistency directly from language. Trained on PartRel3D, DreamPartGen surpasses prior text-to-3D and part-aware baselines, achieving substantial improvements in geometric fidelity ($\downarrow$60% CD, $\downarrow$41% EMD) and text–shape alignment ($\geq \uparrow$20% CLIP/ULIP). We also evaluate generalization to rare parts and held-out relation predicates, improving over prior part-based baselines (14.7–16.3% $\downarrow$ r-FID, 68.2–71.2% $\downarrow$ CD, 39.6–47.9% $\uparrow$ ULIP-T). In summary, our contributions are:

- We introduce **DreamPartGen**, a language-grounded collaborative diffusion framework that unifies geometric, visual, and relational reasoning for coherent and interpretable part-level text-to-3D synthesis.

- We propose **DPLs** and **RSLs** as complementary representations that jointly encode part geometry, appearance, and inter-part relations and are refined together via synchronized co-denoising.
- We curate **PartRel3D**, a large-scale relational dataset with 300K canonicalized functional and spatial triplets for explicit language-based supervision of inter-part relations across 175 object categories.
- DreamPartGen achieves state-of-the-art performance across four benchmarks, improving geometric fidelity by $\downarrow$60% CD / $\downarrow$41% EMD and text–shape alignment by $\geq \uparrow$20% CLIP/ULIP, and showing the same representation supports relational part editing, articulated generation, and mini-scene synthesis.

2 Related Work

Text-to-3D Generation. Early text-to-3D approaches such as DreamFusion [35], ProlificDreamer [45], and LucidDreamer [23] leverage the idea of score distillation sampling (SDS) to generate 3D assets from 2D diffusion priors. While effective for producing single objects, SDS approaches often suffer from low fidelity and poor multi-view consistency [6, 14, 21, 24, 37, 40]. Recent works improve training stability and geometric realism by incorporating differentiable rendering with explicit 3D representations, including Gaussian splatting in DreamGaussian [42] and GaussianDreamer [51], voxel- or mesh-based parameterizations in Clay [55], and hybrid autoregressive architectures such as Trellis [31, 46, 52]. These advances establish strong foundations for high-quality 3D generation, but typically focus on whole objects, without modeling explicit part structure or relational semantics.

Part-level 3D Generation. To address the limitations of object generation, several methods introduce part-aware modeling [5, 10, 13, 20, 25, 49, 50]. Part123 [26] and Salad [18] focus on part segmentation and assembly, while PartGen [4] leverages part decomposition for generative modeling. CoPart [11] extends diffusion models with dual priors over part-level 2D and 3D latents, enabling cross-modality and cross-part mutual guidance. Additionally, works such as PartGS [12] and Part^{2}GS [53] adapt Gaussian splatting for articulated part-aware generation, demonstrating that part supervision yields controllable and physically plausible synthesis [29, 30, 39]. Despite these advances, prior part-aware generation methods rely heavily on geometric signals such as bounding boxes [56], leaving language guidance underexplored, whereas a related line of work uses vision–language models for part-level perception rather than generation [16, 22, 27, 34, 44]. DreamPartGen introduces explicit relational semantic signals that persist throughout denoising, providing both fine-grained part refinement and relation-aware global planning cues directly from natural language.

3 DreamPartGen Method

While recent part-level formulations improve local shape and texture modeling [4, 7, 48], they primarily focus on representation quality and do not explicitly

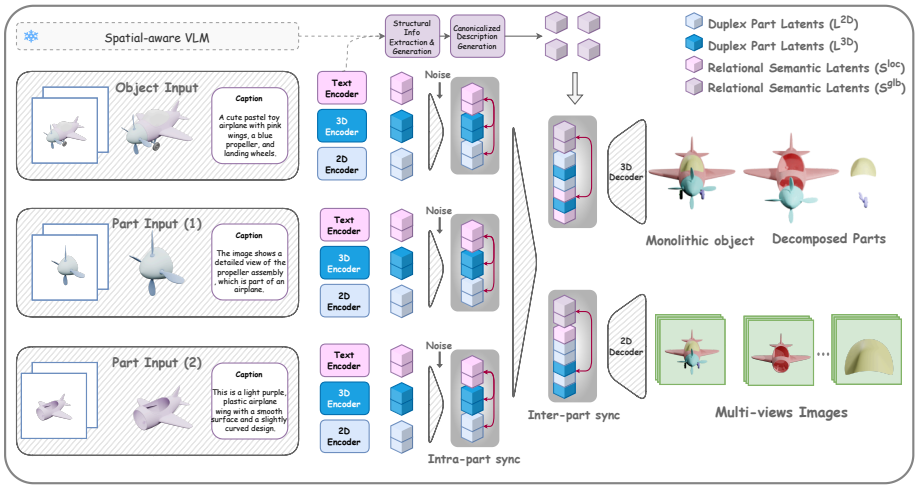

Fig. 2: DreamPartGen Overview. DreamPartGen performs text-guided 3D generation by jointly denoising geometry, appearance, and relational semantics. Each object is decomposed into parts represented as **Duplex Part Latents (DPLs)** from 3D and 2D encoders, while **Relational Semantic Latents (RSLs)** encode text-derived details and global structure. Through intra-part (geometry–appearance alignment) and inter-part (relational planning via language) synchronization, DreamPartGen co-denoises DPLs and RSLs to generate coherent, semantically grounded, part-aware 3D objects.

preserve *text-derived semantics* throughout denoising, which limits their text-to-3D capability and fine-grained controllability. Our key novelty is to introduce *persistent, language-derived relational semantic latents* that remain active throughout the denoising process, rather than using text only as a one-shot condition, and to synchronize them with part-level geometric latents. To this end, we formulate part-based 3D generation as a semantically grounded collaborative diffusion process between two complementary latent representations: ❶ **Duplex Part Latents (DPLs)** (Sec. 3.1), which encode geometry and appearance of individual parts in a modular and disentangled manner, and ❷ **Relational Semantic Latents (RSLs)** (Sec. 3.2), a compact set of text-derived latent tokens that provide both local refinements and global planning signals. During denoising, DPLs and RSLs are synchronized through intra-part and inter-part attention (Sec. 3.3), enabling consistent part-level geometry-appearance alignment and language-guided part assembly. Figure 2 presents an overview of the method.

3.1 Duplex Part Latents (DPLs)

The design of Duplex Part Latents (DPLs) is motivated by recent advances in structured latent representations for 3D generation, which demonstrate that compact latent sets can effectively encode both geometry and appearance [41, 46, 50]. However, existing unified latents primarily operate on voxel-aligned local features capturing shape and texture, but remain tied to spatial grids rather

than semantic components. As a result, they lack modularity across objects and do not support explicit part-level disentanglement or relational reasoning. To address these limitations, we represent each object as a collection of N semantic parts $O = \{p_i\}_{i=1}^{N}$, and encode each part using three complementary elements:

- **3D tokens:** For each part mesh p_i, we sample surface points with associated normals and pass them through a 3D VAE encoder [17, 54], producing a latent sequence $\mathbf{L}_i^{\text{3D}} \in \mathbb{R}^{T_{\text{3D}} \times d}$, where T_{3D} denotes the number of 3D latent tokens and d their embedding dimension, capturing local geometry and spatial structure.
- **2D tokens:** Each part is also rendered from multiple viewpoints, and the resulting images are passed through a pretrained image VAE [3], yielding $\mathbf{L}_i^{\text{2D}} \in \mathbb{R}^{T_{\text{2D}} \times d}$, which encodes color, texture, and shading cues.
- **Part-identity:** To stabilize part tracking across denoising steps, we assign a learnable identifier embedding $e_i \in \mathbb{R}^d$ to each part. These identifiers act as persistent slot identities, binding each latent to its corresponding part and preventing slot swapping across denoising, while relational reasoning layers flexibly reorganize cross-part interactions.

Compared to prior structured latent designs [25, 46], Duplex Part Latents (DPLs) are designed to preserve semantic independence while enabling language-conditioned relational reasoning. This yields several key benefits. First, the architecture is permutation-robust to the input ordering of parts, as the learnable part-identity embeddings prevent semantics from depending on the input part order. Second, the identifiers provide *slot persistence* across denoising timesteps, improving stability of intra-part and inter-part synchronization. Third, because each part is represented as its own modular latent triplet $(\mathbf{L}_i^{\text{3D}}, \mathbf{L}_i^{\text{2D}}, e_i)$, DPLs naturally support cross-object generalization, enabling latent transfer between objects with shared functional components. Finally, DPLs are lightweight and modular, making them directly suitable for integration with the diffusion process, facilitating coherent multi-part synthesis and reasoning.

3.2 Relational Semantic Latents (RSLs)

While DPLs provide modular and disentangled representations for individual parts, they do not by themselves guarantee that the assembled object is globally coherent. This reflects a broader challenge in part-based 3D generation: local geometry and appearance can be faithfully synthesized, yet without explicit semantic coordination, the resulting object structure may violate plausible spatial or functional relations [6, 24, 45]. To address this gap, we introduce Relational Semantic Latents (RSLs), a compact set of *language-derived latent tokens* that provide semantic control signals for part interactions through two roles: persistent global planners and diffused local refiners. In particular, global relational tokens $\mathbf{S}^{\text{glb}}$ persist as fixed structural conditions, while local semantic tokens $\mathbf{S}^{\text{loc}}$ are diffused and denoised alongside the part latents to refine part-level details.

Global Relational Tokens. At the object level, we extract relational phrases from whole-object and part-level descriptions (*e.g.*, "the seat is above the legs," "the propeller is attached to the fuselage," "the two wings are symmetric", *etc.*).

Each phrase is canonicalized into a triplet (i, j, ρ), where i and j denote parts and ρ is a relation predicate such as `support`, `attach`, `symmetry`, or `articulation`. These triplets are assembled into a relational graph and projected into the latent space (see Fig. 3) to yield a set of global relational tokens:

$$\mathbf{S}^{\text{glb}} = \{\mathbf{s}^{\text{glb}}_{ij,\rho}\}_{(i,j,\rho)\in\mathcal{R}}, \quad \mathbf{s}^{\text{glb}}_{ij,\rho} \in \mathbb{R}^d. \tag{1}$$

In this way, $\mathbf{S}^{\text{glb}}$ constitutes a relational graph latent, where each token encodes how two parts are semantically related. These tokens persist throughout the diffusion process and are injected into object-level synchronization, functioning both as semantic planners that specify inter-part relations and as structural conditions that enforce coherent assembly. Unlike prior geometry-based approaches, they are derived from natural language, embedding functional and structural priors without explicit geometric supervision.

Local Semantic Tokens. At the part level, we encode fine-grained semantic cues (*e.g.*, "metallic blade," "wooden handle", *etc.*) to refine material and appearance. Each phrase is encoded and projected into the latent space to yield K_m local semantic tokens:

$$\mathbf{S}^{\text{loc}} = \{\mathbf{s}^{\text{loc}}_{m}\}_{m=1}^{K_m}, \quad \mathbf{s}^{\text{loc}}_{m} \in \mathbb{R}^d, \tag{2}$$

which directly interact with the structural DPL tokens to enhance geometric fidelity and appearance under semantic constraints. Compared to geometry-only latents, RSLs are compact, interpretable, and flexible: their number adapts to object complexity, and additional tokens can be easily obtained by generating short textual descriptions for new parts or relations. Unlike one-shot text conditioning [24, 35], we inject these tokens at every denoising step, enabling iterative semantic refinement. During diffusion, we apply the standard forward noising process to obtain $\mathbf{S}^{\text{loc},t}$ from the clean tokens $\mathbf{S}^{\text{loc}}$. The noised local semantic tokens $\mathbf{S}^{\text{loc},t}$ are injected at each step t to synchronize with the noised part latents and refine part-specific appearance.

3.3 Semantically-Grounded Part Generation

We first instantiate DPLs by encoding each part mesh p_i into geometry and appearance token sequences $(\mathbf{L}^{\text{3D}}_i, \mathbf{L}^{\text{2D}}_i)$ using the 3D VAE encoder and the pre-trained image VAE encoder introduced in Section 3.1, and tag each part with a learnable identifier e_i. We instantiate RSLs by encoding extracted relational/attribute phrases with a frozen text encoder $\mathcal{E}_{\text{text}}$ [43] followed by a learned projection ϕ_{text}, yielding $(\mathbf{S}^{\text{glb}}, \mathbf{S}^{\text{loc}})$. To enable coherent generation, DPLs and RSLs interact throughout denoising via a two-level synchronization mechanism. Specifically, we perform diffusion over the noised part latents $\{\mathbf{L}^{\text{3D},t}_i, \mathbf{L}^{\text{2D},t}_i\}_{i=1}^{N}$ and the noised local semantic tokens $\mathbf{S}^{\text{loc},t}$, while keeping the global relational tokens $\mathbf{S}^{\text{glb}}$ persistent as fixed structural conditions. At each step t, we first apply *intra-part synchronization* to align geometry and appearance within each part under local semantic guidance, and then apply *inter-part synchronization* to propagate context across parts and enforce global relational constraints.

Intra-Part Synchronization. At diffusion step t, each part p_i is represented by a noised geometry latent sequence $\mathbf{L}_i^{3D,t}$ and a noised appearance latent sequence $\mathbf{L}_i^{2D,t}$. We first synchronize these two streams to maintain intra-part geometry-appearance consistency, and then inject noised local semantic tokens $\mathbf{S}^{\text{loc},t}$ to refine part-specific geometric and visual details according to semantic cues. Formally,

$$\begin{aligned}
\mathbf{L}_i^{3D,t} &\leftarrow \mathbf{L}_i^{3D,t} + \alpha_{3D} \cdot \text{Attn}(\mathbf{L}_i^{3D,t}, \mathbf{L}_i^{2D,t}),\\
\mathbf{L}_i^{2D,t} &\leftarrow \mathbf{L}_i^{2D,t} + \alpha_{2D} \cdot \text{Attn}(\mathbf{L}_i^{2D,t}, \mathbf{L}_i^{3D,t}),\\
\mathbf{L}_i^{3D,t} &\leftarrow \mathbf{L}_i^{3D,t} + \lambda_{3D} \cdot \text{Attn}(\mathbf{L}_i^{3D,t}, \mathbf{S}^{\text{loc},t}),\\
\mathbf{L}_i^{2D,t} &\leftarrow \mathbf{L}_i^{2D,t} + \lambda_{2D} \cdot \text{Attn}(\mathbf{L}_i^{2D,t}, \mathbf{S}^{\text{loc},t}),
\end{aligned} \tag{3}$$

where $\alpha_{3D}, \alpha_{2D}, \lambda_{3D}, \lambda_{2D}$ are learnable fusion coefficients.

Inter-Part Synchronization. After intra-part alignment, we propagate context across parts to encourage globally consistent assembly. This is done in two complementary ways: (i) direct message passing among all part latents to share global context, and (ii) relational guidance from persistent global tokens $\mathbf{S}^{\text{glb}}$ that encode inter-part predicates (*e.g.*, `support`, `attach`, `symmetry`, `articulation`). Finally, we update $\mathbf{S}^{\text{glb}}$ via bottom-up grounding from the current part latents, refining the relational plan based on synthesized geometric and appearance evidence. Concretely,

$$\begin{aligned}
\mathbf{L}_i^{3D,t} &\leftarrow \mathbf{L}_i^{3D,t} + \text{Attn}\big(\mathbf{L}_i^{3D,t}, \{\mathbf{L}_j^{3D,t}\}_{j=1}^{N}\big),\\
\mathbf{L}_i^{2D,t} &\leftarrow \mathbf{L}_i^{2D,t} + \text{Attn}\big(\mathbf{L}_i^{2D,t}, \{\mathbf{L}_j^{2D,t}\}_{j=1}^{N}\big),\\
\mathbf{L}_i^{3D,t} &\leftarrow \mathbf{L}_i^{3D,t} + \beta_{3D} \cdot \text{Attn}(\mathbf{L}_i^{3D,t}, \mathbf{S}^{\text{glb}}),\\
\mathbf{L}_i^{2D,t} &\leftarrow \mathbf{L}_i^{2D,t} + \beta_{2D} \cdot \text{Attn}(\mathbf{L}_i^{2D,t}, \mathbf{S}^{\text{glb}}),\\
\mathbf{S}^{\text{glb}} &\leftarrow \mathbf{S}^{\text{glb}} + \eta \cdot \text{Attn}\Big(\mathbf{S}^{\text{glb}}, \{\text{Pool}(\mathbf{L}_i^{3D,t}, \mathbf{L}_i^{2D,t})\}_{i=1}^{N}\Big).
\end{aligned} \tag{4}$$

where $\text{Pool}(\cdot)$ aggregates each part's latent sequences into a compact summary for bottom-up grounding. Note that $\mathbf{S}^{\text{glb}}$ is not a diffusion variable, but rather is updated deterministically as a planner state that remains available as a fixed relational condition at every timestep.

Optimization. Training proceeds in two phases. In the first phase, we optimize diffusion objectives for both 3D and 2D DPLs under semantic conditioning from RSLs. For timesteps $t \sim \mathcal{U}\{1, \ldots, T\}$ and noise $\varepsilon \sim \mathcal{N}(0, I)$, the per-part diffusion losses are:

$$\begin{aligned}
\mathcal{L}_{\text{diff}}^{3D} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{t,\varepsilon}\left[\left\|\varepsilon - \mathcal{N}_{3D}\left(\mathbf{L}_i^{3D,t}, \mathbf{L}_i^{2D,t}, \mathbf{S}^{\text{glb}}, \mathbf{S}^{\text{loc},t}, t\right)\right\|_2^2\right],\\
\mathcal{L}_{\text{diff}}^{2D} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{t,\varepsilon}\left[\left\|\varepsilon - \mathcal{N}_{2D}\left(\mathbf{L}_i^{2D,t}, \mathbf{L}_i^{3D,t}, \mathbf{S}^{\text{glb}}, \mathbf{S}^{\text{loc},t}, t\right)\right\|_2^2\right].
\end{aligned} \tag{5}$$

In the second phase, we fine-tune the model jointly across the 3D and 2D part denoisers and synchronization modules, using an SNR-based weighting scheme.

The overall objective is

$$\mathcal{L} \;=\; \mathbb{E}_t\Big[w_{\text{syn}}(t)\,\big(\mathcal{L}^{\text{3D}}_{\text{diff}} + \mathcal{L}^{\text{2D}}_{\text{diff}}\big)\Big], \tag{6}$$

where weights follow an SNR-based schedule $w_{\text{syn}}(t) = \frac{\text{SNR}(t)}{1+\text{SNR}(t)}$, with $\text{SNR}(t) = \alpha_t^2/\sigma_t^2$ defined by the diffusion coefficients.

Inference. At test time, we encode the input prompt into local semantic tokens $\mathbf{S}^{\text{loc}}$ and global planner tokens $\mathbf{S}^{\text{glb}}$. When explicit triplets are available (*e.g.*, provided by the user or an external parser), they are encoded as $\mathbf{S}^{\text{glb}}$; otherwise, we default to prompt-only conditioning instead of relying on external VLMs to supplement the triplets. This makes the VLM an optional semantic frontend for convenient relation extraction, rather than a required component of the generative backbone. We then initialize part latents by sampling Gaussian noise for the geometry and appearance streams, $\{\mathbf{L}_i^{\text{3D},T}, \mathbf{L}_i^{\text{2D},T}\}_{i=1}^N$, and initialize the local semantic stream by applying the same forward noising process used in training to obtain $\mathbf{S}^{\text{loc},T}$ from $\mathbf{S}^{\text{loc}}$. From timestep T to 1, we jointly denoise $\{\mathbf{L}_i^{\text{3D},t}, \mathbf{L}_i^{\text{2D},t}\}_{i=1}^N$ and $\mathbf{S}^{\text{loc},t}$ using the same part-level and object-level synchronization modules while conditioning on persistent $\mathbf{S}^{\text{glb}}$. After denoising, the final geometry latents $\{\mathbf{L}_i^{\text{3D},0}\}_{i=1}^N$ are decoded by the 3D VAE decoder to obtain part meshes, and the object is assembled from the decoded parts, with $\mathbf{L}_i^{\text{2D},0}$ used for appearance rendering when needed.

4 PartRel3D Dataset

Existing 3D datasets such as PartNet [33], Objaverse [9], and PartVerse [11] provide large-scale geometric diversity but are limited in semantic grounding and relational coverage. They often include either geometry-only annotations or unconstrained text captions without consistent part correspondence, limiting their suitability for training models that understand object assembly (how parts connect) and semantics (what roles parts play). To overcome these limitations, we introduce PartRel3D, a large-scale, relationally annotated extension of PartVerse [11] that links part geometry, appearance, and language through explicit functional and spatial relationships. Each object in PartRel3D is augmented with canonicalized triplets that encode both functional dependencies (*e.g.*, support, attach, hinge) and spatial arrangements (*e.g.*, above, touching, aligned-with), providing large-scale supervision of assembly-level semantics in 3D (Figure 3).

Functional Triplets capture how parts interact in terms of support, attachment, and articulation. Given part- and object-level descriptions, we canonicalize phrases such as "legs support seat" or "handle attached to body" into triplets $(i, j, \rho^{\text{function}})$, where i, j are part indices and ρ^{function} is a functional predicate (*e.g.*, `support`, `attach`, `hinge`, `symmetry`).

Spatial Triplets capture geometric and positional relations between parts. Each triplet has the same form $(i, j, \rho^{\text{spatial}})$, where i and j still index parts from the set of object parts $\mathcal{P}$, but ρ^{spatial} is a predicate drawn from a controlled vocabulary of interpretable, assembly-relevant predicates. These include vertical relations

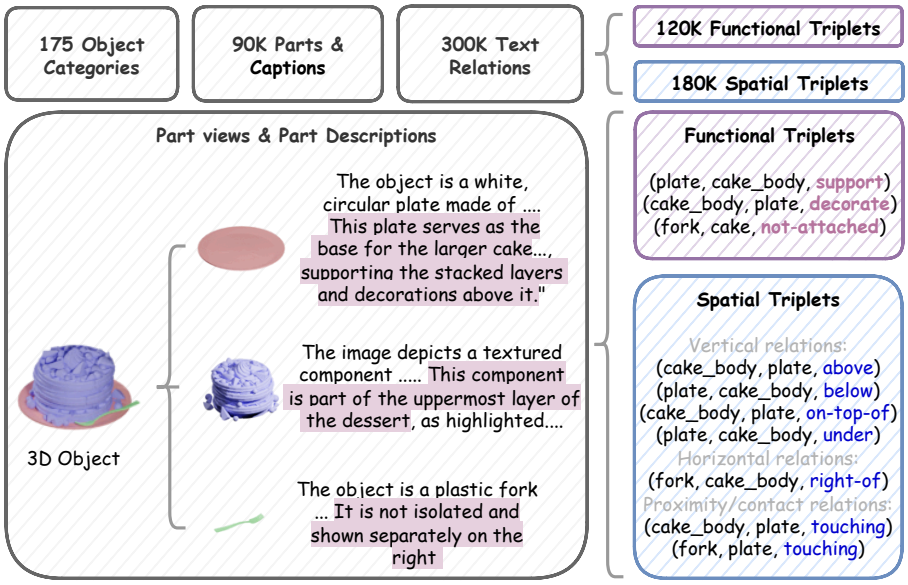

Fig. 3: PartRel3D dataset overview of structured functional and spatial triplets for fine-grained inter-part semantic supervision.

(`above`, `below`, `on-top-of`, `under`), horizontal relations (`in-front-of`, `behind`, `left-of`, `right-of`), containment relations (`inside`, `surrounding`), symmetry/arrangement (`symmetric-with`, `parallel-to`, `aligned-with`), proximity and contact relations (`touching`, `attached-to`, `connected-with`).

The resulting PartRel3D dataset contains approximately 11K part-labeled objects spanning 175 object categories, with over 90K individual parts and 300K canonicalized relational triplets. On average, each object contains 8.2 parts and 27 inter-part relations, providing dense structural supervision. Additional details, dataset statistics, and canonicalization criteria are available in the Appendix A.

5 Experiments

Baselines. We compare DreamPartGen against Trellis [46], CLAY [55], HoloPart [49], and PartCrafter [25], as they represent the current state of the art in 3D generation. These methods collectively capture the diversity of contemporary approaches from structured latent representations [46] to explicit part generation and assembly part-aware text-driven 3D [25]. Moreover, all of them provide open-source implementations, enabling fair and reproducible comparison under consistent training and evaluation protocols. Additional details in Appendix B.
Metrics. Following prior work [55], we adopt both perceptual and structural metrics for text-to-3D evaluation. We report render-FID and render-KID (r-FID/r-KID) computed from multi-view renderings to assess visual fidelity, and P-FID/P-KID computed in 3D feature space using PointNet++ [36]. Chamfer Distance (CD)

Table 1: Quantitative evaluation on 3D object generation. Quantitative comparison with state-of-the-art methods on Objaverse, ShapeNet, ABO, and PartRel3D. Highlighted best and second-best results.

Method	Objaverse			ShapeNet			ABO			PartRel3D		
	CD↓	EMD↓	IoU↓	CD↓	EMD↓	IoU↓	CD↓	EMD↓	IoU↓	CD↓	EMD↓	IoU↓
Trellis	0.361	1.320	-	0.549	1.482	-	0.287	0.933	-	0.532	1.526	-
CLAY	0.318	1.245	-	0.527	1.503	-	0.321	1.022	-	0.410	1.646	-
HoloPart	0.334	1.298	0.494	0.478	1.354	0.542	0.269	0.911	0.529	0.355	1.623	0.716
PartCrafter	0.278	1.107	0.453	0.451	1.252	0.499	0.266	0.905	0.505	0.371	1.474	0.700
DreamPartGen	0.141	0.810	0.359	0.222	0.967	0.503	0.101	0.531	0.404	0.081	0.412	0.304

and Earth Mover's Distance (EMD) measure geometric precision. For text–shape alignment, we compute similarity with CLIP-ViT/L-14 [38] and ULIP [47]. In particular, ULIP-T is defined as the inner product between normalized ULIP embeddings of the caption T and the generated shape S, ULIP-T$(T, S) = \langle E_T, E_S \rangle$, reflecting the semantic coherence between textual and geometric modalities. We further use the average pairwise Intersection-over-Union (IoU) to evaluate the geometric independence of generated part meshes. Specifically, we voxelize each generated part in a shared canonical space using a $64 \times 64 \times 64$ grid and compute the average pairwise IoU across all generated parts following [25]. Lower IoU indicates less inter-part overlap and therefore better part disentanglement. The ideal case is that generated parts are non-intersecting while remaining composable into a plausible object consistent with the ground-truth structure.

5.1 Quantitative Results

We evaluate geometric reconstruction quality across Objaverse [9], ShapeNet [2], ABO [8], and PartRel3D. As shown in Table 1, DreamPartGen consistently achieves the lowest CD and EMD on all benchmarks, outperforming prior methods by large margins (↓60% CD and ↓41% EMD on average). Moreover, DreamPartGen attains the lowest IoU scores (↓ 27.2% on average relative to the strongest baseline across each benchmark), reflecting stronger *geometry independence*, *i.e.*, the ability to generate non-intersecting yet composable parts that maintain object-level coherence.

We further assess text–shape alignment on the PartVerse dataset following [11], where half of the test cases describe individual parts (*e.g.*, "a chair leg") and the rest correspond to complete objects. Table 2 shows DreamPartGen improves text-shape alignment over the strongest baseline across all metrics by ($\geq$ 20%) at the object

Table 2: Text-shape alignment comparison. Quantitative comparison on PartVerse. Highlighted best and second-best.

Scope	Method	CLIP(N-T)↑	CLIP(I-T)↑	ULIP-T↑
Object-level	Trellis	0.192	0.214	0.164
	CLAY	0.194	0.216	0.156
	HoloPart	0.186	0.206	0.155
	PartCrafter	0.187	0.207	0.162
	DreamPartGen	0.235	0.264	0.197
Part-level	Trellis	0.106	0.122	0.091
	CLAY	0.112	0.128	0.096
	HoloPart	0.130	0.141	0.113
	PartCrafter	0.125	0.145	0.109
	DreamPartGen	0.179	0.200	0.153

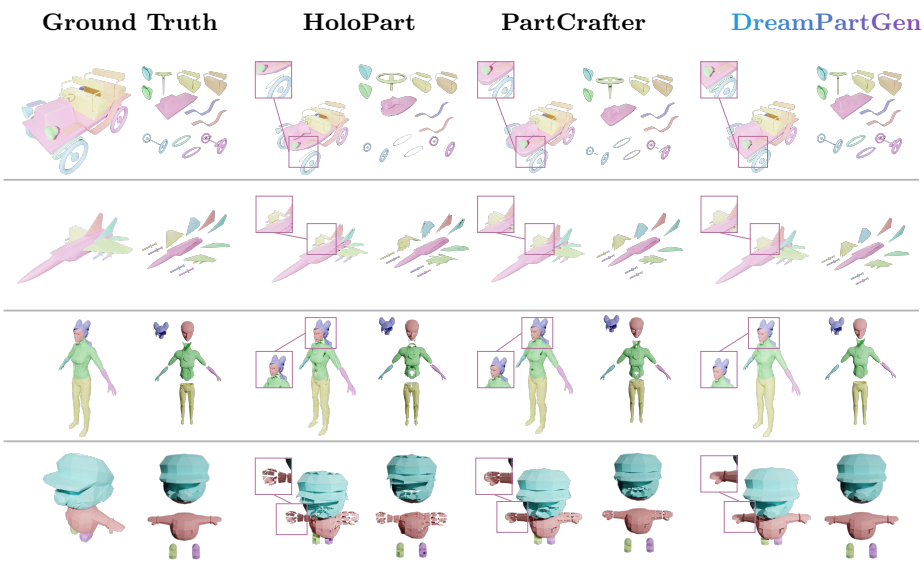

Fig. 4: Qualitative comparison on part-level 3D generation. Across diverse object categories, DreamPartGen yields the most faithful decompositions, preserving clear part boundaries, correct topology, and consistent spatial alignment. Baselines frequently exhibit assembly failures such as missing or detached parts (*e.g.*, wing/head), spatial drift of small components (*e.g.*, wheels/mechanical parts floating off the chassis), and unstable attachments that create surface tearing or holes around high-contact regions (neck, torso, shoulders, limb joints).

level and ($\geq 35\%$) at the part level, highlighting the effectiveness of RSLs for fine-grained semantic grounding.

5.2 Qualitative Results

Figure 4 highlights that, across diverse object categories, DreamPartGen consistently generates 3D objects with part-consistent and physically plausible assemblies. Compared to the strongest baselines, HoloPart [49] and PartCrafter [25], DreamPartGen preserves fine-grained geometry more faithfully, maintains inter-part relationships better, and respects global structural constraints that are frequently violated by prior approaches. As illustrated, baselines frequently omit distorted parts or misplace them in space, for instance, generating wheels that float away from the chassis or misaligning small mechanical parts, leading to broken functional geometry in the first example. Similar failures appear in the second and third examples, where HoloPart produces a detached wing (airplane) or head (humanoid), and both baselines exhibit surface tearing and holes around the neck, torso, and shoulders, indicating incomplete and unstable attachment geometry. DreamPartGen, by contrast, generates watertight meshes with intact intra-part connections, smoother surfaces, and correctly integrated parts. Finally, in the last row, baselines suffer from hollow torsos, shredded hand geometry, and

Fig. 5: Ablation on the co-denoising process with local semantic tokens $\mathbf{S}^{\mathbf{loc}}$. The condition-only baseline ($-\mathbf{S}^{\mathrm{loc}}$) yields coarse geometry and weak semantic coherence between parts.

Fig. 6: Ablation on the global relational tokens $\mathbf{S}^{\mathbf{glb}}$. The model exhibits part-level misalignment and spatial drift without $\mathbf{S}^{\mathrm{glb}}$.

broken limb attachments, while DreamPartGen maintains coherent small-part geometry and avoids the severe tearing and disintegration observed in prior methods. Together, these results demonstrate that DreamPartGen's relationally grounded generation maintains local part fidelity but also enforces globally consistent part connectivity even in complex articulated 3D structures.

5.3 Ablations

Relational Semantic Latents (RSLs) Qualitative Analysis. We qualitatively analyze the roles of the local semantic $\mathbf{S}^{\mathrm{loc}}$ and global relational tokens $\mathbf{S}^{\mathrm{glb}}$. For $\mathbf{S}^{\mathrm{loc}}$, we compare against a conditioning-only baseline where text embeddings are injected only via timestep-wise cross-attention, without maintaining persistent semantic latents across denoising. As shown in Figure 5, conditioning-only yields coarser and less consistent surface geometry, with weaker semantic coherence between parts, indicating that co-denoising with $\mathbf{S}^{\mathrm{loc}}$ is essential for high-fidelity part synthesis and stable semantic consistency throughout generation. For $\mathbf{S}^{\mathrm{glb}}$, we remove the persistent global relational tokens and their object-level synchronization while keeping part-level denoising unchanged. Without global relational guidance, parts remain plausible in isolation but the assembled object exhibits increased inter-part misalignment, weaker structural coherence, and spatial drift (Figure 6), confirming that global relational semantics are crucial for enforcing coherent object-level organization during denoising.

RSLs and Part Identity Quantitative Analysis. Table 3 summarizes the contribution of each component in DreamPartGen, evaluated on a test subset of PartRel3D dataset

Table 3: Ablation on Model Components. Evaluation on a fixed random PartRel3D subset. Best and second-best results are highlighted.

Method	CD↓	EMD↓	IoU↓	ULIP-T↑
HoloPart	0.226	1.482	0.318	0.112
PartCrafter	0.219	1.403	0.341	0.101
DreamPartGen	0.145	0.771	0.212	0.158
✗ $\mathbf{S}^{\mathrm{glb}}$	0.292	2.892	0.587	0.084
✗ $\mathbf{S}^{\mathrm{loc}}$	0.781	5.764	0.652	0.089
✗ **Part Identifier**	0.277	1.709	0.438	0.091

Table 4: Inference-time Comparison. We report per-sample inference latency. Methods are grouped by task type (object-level, part-level, or scene-level generation); timings should be interpreted within each row. Best is highlighted.

Task Setting	HoloPart	PartCrafter	TRELLIS	CLAY	MIDI	DreamPartGen
Object Generation	–	–	95s	118s	–	45s
Part-level Generation	21m	112s	–	–	–	109s
3D Scene Generation	–	64s	–	–	102s	52s

and compared against strong part-aware baselines. We assess performance using geometric fidelity (CD, EMD), part-level separation via pairwise IoU, and text–shape alignment through ULIP-T, resulting in a comprehensive view of both geometry and semantics. Removing the global relational tokens (✗ $\mathbf{S}^{\text{glb}}$) causes a major regression: CD increases from 0.145 to 0.292 (↑101.4%), EMD increases from 0.771 to 2.892 (↑275.1%), and part overlap rises from 0.212 to 0.587 (↑176.9%), while ULIP-T drops from 0.158 to 0.084 (↓46.8%), indicating that relational context is essential for preventing collisions and maintaining coherent assembly. Disabling the local semantic tokens (✗ $\mathbf{S}^{\text{loc}}$) further degrades performance: CD increases to 0.781 (↑ 438.6%), EMD increases to 5.764 (↑ 647.6%), IoU increases to 0.652 (↑ 207.5%), and ULIP-T decreases to 0.089 (↓ 43.7%), confirming the importance of jointly evolving part and semantic latents for stable generation. Eliminating the part identifier module (✗ **Part Identifier**) also hurts disentanglement and semantics: IoU increases to 0.438 (↑106.6%), CD/EMD increase to 0.277/1.709 (↑ 91.0%/ ↑ 121.7%), and ULIP-T drops to 0.091 (↓42.4%), showing it helps preserve identity-consistent structure.

Inference Efficiency. Table 4 compares per-sample inference latency across representative 3D generation methods. Since these methods target different generation settings (object-level, part-level, and scene-level), we group comparisons by task type and interpret timings within each row. For DreamPartGen, we report the *prompt-only* setting to isolate the cost of the generative backbone. Results show DreamPartGen remains efficient despite its semantic synchronization design.

Additional Ablations. Appendix C reports additional experiments, including perceptual metrics, part-level reconstruction quality, ablations on the number of RSL tokens, and robustness to relation parsing, rare parts, and held-out relations.

5.4 Downstream Applications

Text-to-3D Scene Generation. DreamPartGen enables a wide range of part-aware 3D applications, including text-to-3D scene generation. In this task, the goal is to generate a coherent multi-object scene (a small scene) directly from a text prompt. During generation, each object is treated as a macro-part with aggregated DPLs, and a scene-level relational graph from canonicalized triplets (o_i, o_j, ρ) encodes spatial and functional relations. Objects are first generated independently and then jointly refined in a brief synchronization step to produce the final, coherent scene. Additional details on the scene generation process can

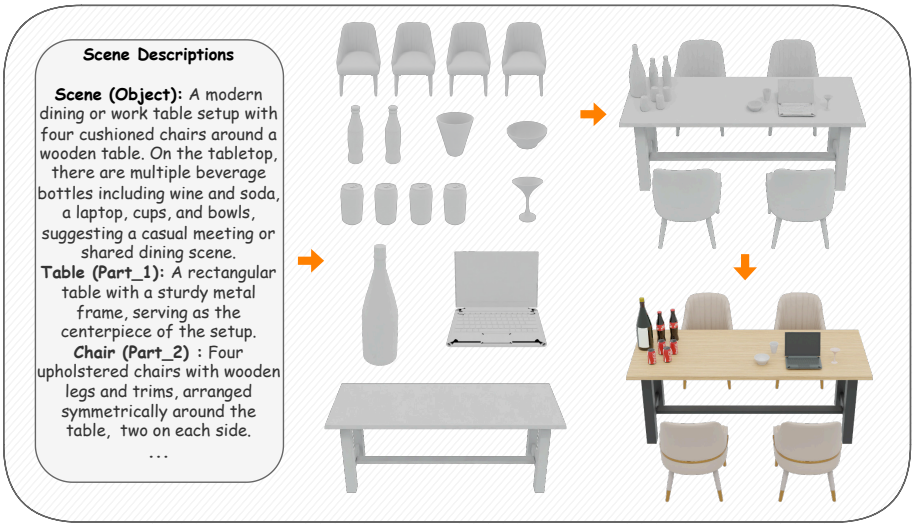

Fig. 7: Mini-scene generation. Given a scene-level description, DreamPartGen can generate a complete, coherent 3D scene with physically plausible spatial layouts, capturing object geometry and fine-grained object- and part-level relations.

be found in Appendix D. As shown in Figure 7, DreamPartGen can synthesize multi-object scenes that respect part structure, spatial constraints, and global coherence. DreamPartGen's DPLs assign persistent, semantically meaningful slots for every part category, ensuring that the model explicitly reasons over fine-grained sub-components (*e.g.*, wooden chair legs) and part counts (*e.g.*, four chairs). Additional examples are provided in Figure 1 (bottom right) and in Appendix D.

Text-to-3D Part Editing. To edit a specific part, we isolate its DPLs and freeze all others while keeping the global relational context fixed. We then apply localized redenoising via partial DDIM inversion, optimizing only the target part's latents, followed by a brief synchronization step to restore coherence with the full object. As illustrated in Figure 8, DreamPartGen accurately executes relational part editing prompts, producing clean, high-fidelity edits with seamless part-to-part coherence. An additional editing example is shown in Figure 1 (top right). Details and additional qualitative examples can be found in Appendix D.

Fig. 8: Part-level editing. From a source object (left), DreamPartGen can correctly execute relational edit prompts to place accessories on top of or around the head, while preserving geometry and spatial consistency.

6 Conclusion

We introduce DreamPartGen, a part-aware text-to-3D generation framework that bridges geometric structure and semantic reasoning through collaborative part-latent denoising. By coupling Duplex Part Latents (DPLs) with Relational Semantic Latents (RSLs), DreamPartGen jointly models geometry, appearance, and inter-part relations, enabling coherent, interpretable, and controllable 3D synthesis. Beyond single-object generation, DreamPartGen enables a broad suite of part-centric applications, including relational part editing and compositional scene generation, highlighting the benefits of explicitly modeling 3D objects through structured, semantically grounded part latents. An exciting future direction is to extend this formulation to densely entangled structures, such as bushes and other highly intricate objects or scenes.

Acknowledgments

This research was partially supported by Google, the Google TPU Research Cloud (TRC) program, the Lambda Research Grant, the NSF CAREER Award #2542328, and the U.S. Army under contract W5170125CA160. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Google, Lambda, NSF, the U.S. Army, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 1
2. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 10
3. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: International Conference on Learning Representations (ICLR). vol. 2024, pp. 57611–57640 (2024) 5
4. Chen, M., Shapovalov, R., Laina, I., Monnier, T., Wang, J., Novotny, D., Vedaldi, A.: Partgen: Part-level 3d generation and reconstruction with multi-view diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5881–5892 (2025) 2, 3
5. Chen, M., Wang, J., Shapovalov, R., Monnier, T., Jung, H., Wang, D., Ranjan, R., Laina, I., Vedaldi, A.: Autopartgen: Autoregressive 3d part generation and discovery. Advances in Neural Information Processing Systems (NeurIPS) pp. 153496–153521 (2026) 3

6. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: International Conference on Computer Vision (ICCV). pp. 22246–22256 (2023) 2, 3, 5
7. Chen, Y., Li, Z., Wang, Y., Zhang, H., Li, Q., Zhang, C., Lin, G.: Ultra3d: Efficient and high-fidelity 3d generation with part attention. arXiv preprint arXiv:2507.17745 (2025) 3
8. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., et al.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21126–21136 (2022) 10
9. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13142–13153 (2023) 8, 10
10. Ding, L., Dong, S., Li, Y., Gao, C., Chen, X., Han, R., Kuang, Y., Zhang, H., Huang, B., Huang, Z., et al.: Fullpart: Generating each 3d part at full resolution. arXiv preprint arXiv:2510.26140 (2025) 3
11. Dong, S., Ding, L., Chen, X., Li, Y., Wang, Y., Wang, Y., Wang, Q., Kim, J., Gao, C., Huang, Z., et al.: From one to more: Contextual part latents for 3d generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8230–8240 (2025) 3, 8, 10, 1
12. Gao, Z., Yi, R., Huang, Y., Chen, W., Zhu, C., Xu, K.: Partgs: Learning part-aware 3d representations by fusing 2d gaussians and superquadrics. arXiv preprint arXiv:2408.10789 (2024) 2, 3, 7
13. Hertz, A., Perel, O., Giryes, R., Sorkine-Hornung, O., Cohen-Or, D.: Spaghetti: Editing implicit shapes through part aware generation. ACM Transactions on Graphics (TOG) **41**(4), 1–20 (2022) 3
14. Hu, Z., Zhao, M., Zhao, C., Liang, X., Li, L., Zhao, Z., Fan, C., Zhou, X., Yu, X.: Efficientdreamer: High-fidelity and robust 3d creation via orthogonal-view diffusion priors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4949–4958 (2024) 3
15. Huang, Z., Guo, Y.C., An, X., Yang, Y., Li, Y., Zou, Z.X., Liang, D., Liu, X., Cao, Y.P., Sheng, L.: Midi: Multi-instance diffusion for single image to 3d scene generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23646–23657 (2025) 7
16. Kareem, A., Lahoud, J., Cholakkal, H.: Paris3d: Reasoning-based 3d part segmentation using large multimodal model. In: European Conference on Computer Vision (ECCV). pp. 466–482. Springer (2024) 3
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 5
18. Koo, J., Yoo, S., Nguyen, M.H., Sung, M.: Salad: Part-level latent diffusion for 3d shape generation and manipulation. In: International Conference on Computer Vision (ICCV). pp. 14441–14451 (2023) 2, 3
19. Laga, H., Mortara, M., Spagnuolo, M.: Geometry and context for semantic correspondences and functionality recognition in man-made 3d shapes. ACM Transactions on Graphics (TOG) **32**(5), 1–16 (2013) 2
20. Li, S., Paschalidou, D., Guibas, L.: Pasta: Controllable part-aware shape generation with autoregressive transformers. arXiv preprint arXiv:2407.13677 (2024) 3
21. Li, W., Chen, R., Chen, X., Tan, P.: Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. arXiv preprint arXiv:2310.02596 (2023) 3

22. Li, X., Juvekar, A., Zhang, J., Liu, X., Wahed, M., Nguyen, K.A., Shen, Y., Yu, T., Lourentzou, I.: Counterfactual segmentation reasoning: Diagnosing and mitigating pixel-grounding hallucination. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7450–7460 (2026) 3
23. Liang, Y., Yang, X., Lin, J., Li, H., Xu, X., Chen, Y.: Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6517–6526 (2024) 2, 3
24. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 300–309 (2023) 2, 3, 5, 6
25. Lin, Y., Lin, C., Pan, P., Yan, H., Yiqiang, F., Mu, Y., Fragkiadaki, K.: Partcrafter: Structured 3d mesh generation via compositional latent diffusion transformers. Advances in Neural Information Processing Systems (NeurIPS) pp. 35387–35415 (2026) 3, 5, 9, 10, 11
26. Liu, A., Lin, C., Liu, Y., Long, X., Dou, Z., Guo, H.X., Luo, P., Wang, W.: Part123: part-aware 3d reconstruction from a single-view image. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024) 2, 3
27. Liu, M., Zhu, Y., Cai, H., Han, S., Ling, Z., Porikli, F., Su, H.: Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21736–21746 (2023) 3
28. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: International Conference on Computer Vision (ICCV). pp. 9298–9309 (2023) 2
29. Liu, Y., Jia, B., Lu, R., Ni, J., Zhu, S.C., Huang, S.: Building interactable replicas of complex articulated objects via gaussian splatting. In: The Thirteenth International Conference on Learning Representations (2025) 3
30. Lu, R., Liu, Y., Tang, J., Ni, J., Wang, Y., Wan, D., Zeng, G., Chen, Y., Huang, S.: Dreamart: Generating interactable articulated objects from a single image. arXiv preprint arXiv:2507.05763 (2025) 3
31. Lu, S., Lin, H., Yao, L., Gao, Z., Ji, X., Liang, Y., Zhang, L., Ke, G., et al.: Unified cross-scale 3d generation and understanding via autoregressive modeling. arXiv preprint arXiv:2503.16278 (2025) 3
32. Mitra, N.J., Wand, M., Zhang, H., Cohen-Or, D., Kim, V., Huang, Q.X.: Structure-aware shape processing. In: ACM SIGGRAPH 2014 Courses, pp. 1–21. Association for Computing Machinery (2014) 2
33. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 909–918 (2019) 8, 1
34. Nguyen, K.A., Juvekar, A., Yu, T., Wahed, M., Lourentzou, I.: Calico: Part-focused semantic co-segmentation with large vision-language models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4550–4561 (2025) 3
35. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) 2, 3, 6
36. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems (NeurIPS) (2017) 9, 2

37. Qiu, L., Chen, G., Gu, X., Zuo, Q., Xu, M., Wu, Y., Yuan, W., Dong, Z., Bo, L., Han, X.: Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9914–9925 (2024) 3
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021) 10
39. Shen, L., Zhang, S., Li, H., Yang, P., Huang, Z., Zhang, Z., Zhao, H.: Gaussianart: Unified modeling of geometry and motion for articulated objects. In: 2026 International Conference on 3D Vision (3DV). pp. 384–395. IEEE (2026) 3
40. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023) 2, 3
41. Tang, J., Lu, R., Li, M., Hao, Z., Li, X., Wei, F., Song, S., Zeng, G., Liu, M.Y., Lin, T.Y.: Efficient part-level 3d object generation via dual volume packing. Advances in Neural Information Processing Systems (NeurIPS) pp. 27115–27137 (2026) 4
42. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In: International Conference on Learning Representations. vol. 2024, pp. 33879–33896 (2024) 3
43. Team, G., Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al.: Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118 (2024) 6
44. Wahed, M., Nguyen, K.A., Juvekar, A.S., Li, X., Zhou, X., Shah, V., Yu, T., Yanardag, P., Lourentzou, I.: Prima: Multi-image vision-language models for reasoning segmentation. arXiv preprint arXiv:2412.15209 (2024) 3
45. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems (NeurIPS) pp. 8406–8441 (2023) 2, 3, 5
46. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21469–21480 (2025) 3, 4, 5, 9
47. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1179–1189 (2023) 10
48. Yan, X., Xu, J., Li, Y., Ma, C., Yang, Y., Wang, C., Zhao, Z., Lai, Z., Zhao, Y., Chen, Z., et al.: X-part: high fidelity and structure coherent shape decomposition. arXiv preprint arXiv:2509.08643 (2025) 3
49. Yang, Y., Guo, Y.C., Huang, Y., Zou, Z.X., Yu, Z., Li, Y., Cao, Y.P., Liu, X.: Holopart: Generative 3d part amodal segmentation. arXiv preprint arXiv:2504.07943 (2025) 2, 3, 9, 11
50. Yang, Y., Zhou, Y., Guo, Y.C., Zou, Z.X., Huang, Y., Liu, Y.T., Xu, H., Liang, D., Cao, Y.P., Liu, X.: Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025) 3, 4
51. Yi, T., Fang, J., Wang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6796–6807 (2024) 3

52. Yu, T., Li, X., Shen, Y., Liu, Y., Lourentzou, I.: CoRe3D: Collaborative reasoning as a foundation for 3d intelligence. arXiv preprint arXiv:2512.12768 (2025) 3
53. Yu, T., Shah, V., Wahed, M., Shen, Y., Nguyen, K.A., Lourentzou, I.: Part^{2}GS: Part-aware modeling of articulated objects using 3d gaussian splatting. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18913–18923 (2026) 2, 3, 7
54. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. ACM Transactions On Graphics (TOG) **42**(4), 1–16 (2023) 5
55. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. ACM Transactions on Graphics (TOG) **43**(4), 1–20 (2024) 3, 9
56. Zhu, Z., Wan, L., Xu, R., Zhang, Y., Chen, H., Dou, Z., Lin, C., Liu, Y., Wei, M.: Partsam: A scalable promptable part segmentation model trained on native 3d data. arXiv preprint arXiv:2509.21965 (2025) 3