FREEPART: TOWARDS GENERATION-FRIENDLY DECOMPOSITION FOR 3D PART SYNTHESIS

Anonymous authors
Paper under double-blind review

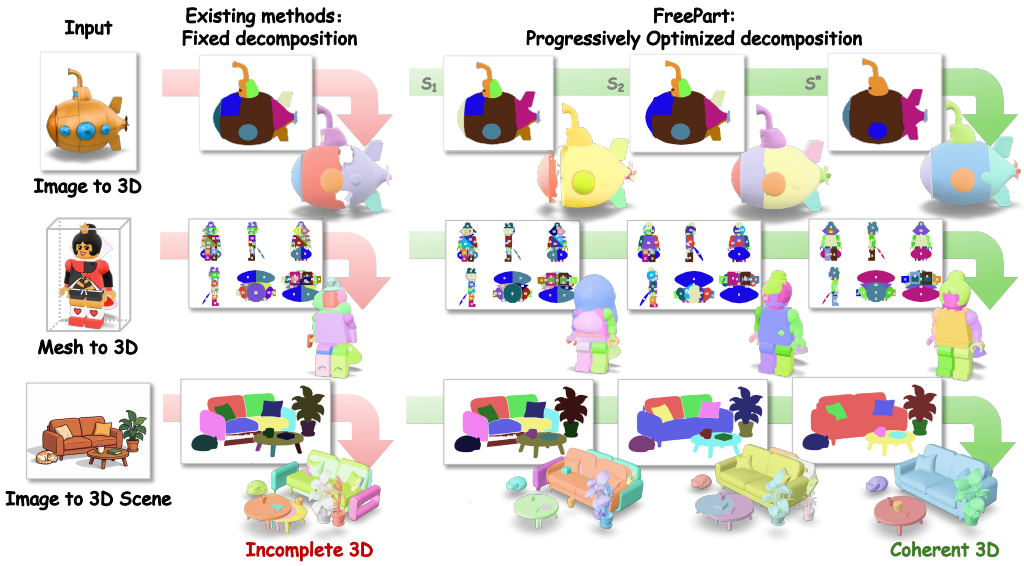


Figure 1: **Motivation and qualitative results of FreePart.** Existing methods rely on fixed decompositions, which may propagate decomposition errors to downstream generation and produce incomplete 3D results. Under the guidance of Part-level Decomposition Assessment (PDA) and Multi-agent Collaborative Search (MCS), FreePart progressively optimizes the decomposition process. Here, $\mathcal{S}_1$ and $\mathcal{S}_2$ denote decomposition states along the search trajectory, while $\mathcal{S}^*$ denotes the final optimized state. This refinement improves part coherence and overall completeness across image-to-3D object, mesh-to-3D object, and image-to-3D scene generation.

ABSTRACT

Part-level 3D generation aims to synthesize complex 3D objects by decomposing them into semantically meaningful parts and generating them separately before assembly. However, we find that even semantically plausible decompositions can lead to substantially different generation difficulties and 3D reconstruction quality. When such decomposition choices are unfavorable, they can introduce errors that propagate through downstream generation, compromising the structural coherence and completeness of the resulting objects. To address this, we propose **FreePart**, a plug-and-play framework that searches for generation-friendly part decompositions and seamlessly integrates them into diverse part-level 3D generation pipelines. Specifically, we first introduce a Part-level Decomposition Assessment (PDA) to evaluate the generation difficulty induced by different decompositions. Guided by PDA, we then design a Multi-agent Collaborative Search (MCS) method, where three complementary agents iteratively diagnose, refine, and evaluate candidate decompositions, with valid decomposition states organized as a directed acyclic graph for efficient search. Extensive experiments across multiple generation pipelines demonstrate that FreePart consistently improves both part-level and overall-object 3D generation quality. Code and website are available in the supplementary material.

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.

1 INTRODUCTION

Recent advances in 3D generation have enabled increasingly realistic and detailed 3D content (Poole et al., 2023; Lin et al., 2023; Wang et al., 2023; Liu et al., 2023; Shi et al., 2024; Hong et al., 2024). Nevertheless, synthesizing complex objects composed of multiple meaningful components remains challenging. Part-level 3D generation provides a structured paradigm for synthesizing complex objects by decomposing them into semantically meaningful parts, generating these parts separately, and assembling them into a complete object. This paradigm enables flexible part-level control and editing while facilitating the generation of complex 3D objects (Liu et al., 2024; Chen et al., 2025a; Lin et al., 2025; Yang et al., 2025; Chen et al., 2025b).

Existing part-level 3D generation methods generally follow a decomposition-then-generation pipeline, where an object is first partitioned into parts and each part is subsequently generated or reconstructed in 3D. Recent works have largely focused on improving the downstream generation stage through stronger generative priors, specialized part representations, or more effective assembly mechanisms (Liu et al., 2024; Chen et al., 2025a; Lin et al., 2025; Yang et al., 2025). In contrast, the initial decomposition is typically treated as a fixed input rather than an optimizable component.

However, we find that a semantically plausible decomposition is not necessarily favorable for 3D generation. Different plausible decompositions of the same object can lead to substantially different generation difficulties and 3D outcomes. As illustrated in Fig. 1 (left), an unfavorable decomposition may make individual parts difficult to reconstruct, eventually resulting in fragmented geometry, missing structures, or incomplete objects. This suggests that part decomposition is not merely a preprocessing choice, but an important factor governing downstream 3D generation quality.

A straightforward solution is to search over different decomposition candidates and select the one that yields the best 3D result. However, this is difficult to achieve in practice. The space of possible decompositions grows rapidly with the number of parts, while evaluating each candidate requires repeatedly running the downstream 3D generation pipeline. Therefore, the key challenge is to identify generation-friendly decompositions without exhaustively generating and comparing all possible candidates.

To address this challenge, we first investigate what makes a part decomposition favorable for downstream generation. Through extensive experiments, we find that the **generation difficulty** of individual parts provides a practical signal for assessing a decomposition: decompositions associated with easier-to-generate parts generally lead to better 3D reconstructions. We further observe that such difficulty can be characterized from two complementary aspects: Surface Complexity (SC) reflects the geometric complexity of a part, while Occlusion Rate (OCC) captures the amount of missing visual evidence for recovering its 3D geometry. Our analysis shows that both factors are closely related to downstream reconstruction quality, providing useful guidance for evaluating and refining part decompositions.

Based on these observations, we propose FreePart, a plug-and-play framework that searches for generation-friendly part decompositions before downstream 3D generation. FreePart first introduces Part-level Decomposition Assessment (PDA), which models part generation difficulty using SC and OCC and provides generation-aware supervision for decomposition optimization. We further develop Multi-agent Collaborative Search (MCS), where three complementary agents iteratively refine the decomposition. The Diagnoser identifies parts that are difficult to generate, the Strategist proposes candidate refinements, and the Evaluator determines whether a proposed refinement should be accepted or rolled back. Together, they form a collaborative diagnosis–proposal–evaluation loop that progressively improves the decomposition toward one better suited for downstream generation.

To efficiently navigate the decomposition space, we organize valid decomposition states as a directed acyclic graph (DAG), where each transition corresponds to a candidate part merge. Rather than exhaustively constructing and evaluating all possible decompositions, MCS selectively explores promising transitions and revisits alternatives when a proposal is rejected. This structured search enables FreePart to progressively optimize the decomposition while preserving semantic and structural validity. Importantly, FreePart does not modify the underlying 3D generator. Once trained, the same framework can be directly integrated into different part-level 3D generation pipelines without generator-specific retraining.

We evaluate FreePart across three representative settings: image-to-3D object generation, mesh-based part generation, and image-to-3D scene generation. Extensive quantitative and qualitative results demonstrate that FreePart consistently improves both part-level and overall 3D generation quality across multiple benchmarks and baseline generators. These results demonstrate that optimizing the decomposition itself provides a general and effective way to improve part-level 3D generation without modifying the underlying generator.

Our contributions are summarized as follows:

- We reveal the critical role of part decomposition in downstream 3D generation and identify Surface Complexity and Occlusion Rate as two key indicators of part generation difficulty.
- We propose **FreePart**, a plug-and-play framework that combines Part-level Decomposition Assessment with Multi-agent Collaborative Search to find generation-friendly decompositions.
- FreePart consistently improves part-level and overall 3D generation quality across diverse generation pipelines.

2 Related Work

General 3D Generation. General 3D generation aims to reconstruct or synthesize complete 3D objects from multimodal conditions. Early methods distilled pretrained 2D diffusion priors into 3D representations via per-instance optimization. DreamFusion introduced Score Distillation Sampling, while Magic3D and ProlificDreamer improved generation quality through coarse-to-fine optimization and Variational Score Distillation (Poole et al., 2023; Lin et al., 2023; Wang et al., 2023). However, these methods remain computationally expensive and susceptible to cross-view inconsistency and Janus artifacts. Subsequent approaches improved consistency and efficiency through viewpoint-conditioned or multi-view diffusion (Liu et al., 2023; Shi et al., 2024) and feed-forward reconstruction with large-scale Transformers (Hong et al., 2024). Recent methods further enhance scalability, geometric fidelity, and structural reliability using structured 3D latents, explicit point priors, native 3D representations, and 3D-aware reasoning (Xiang et al., 2025; Huang et al., 2025b; Wen et al., 2025; Xiang et al., 2026; Chen et al., 2026; Wang et al., 2026; Liu et al., 2026b). Despite these advances, most general 3D generation methods still model an object primarily as a whole, with limited consideration of how it should be organized into generation-friendly local units. Our work instead focuses on generation-oriented part decomposition and optimizes the initial partition before 3D generation.

Part-Level 3D Generation. Part-level 3D generation decomposes complex objects into local units and reconstructs or generates them separately. Part123 introduces multi-view 2D segmentation into single-image reconstruction (Liu et al., 2024); PartGen adopts a "decompose–complete–reconstruct" paradigm for view-consistent part reconstruction (Chen et al., 2025a); and HoloPart extends part modeling to 3D amodal segmentation by recovering complete geometry from partial observations (Yang et al., 2026b). Recent methods further develop dedicated part-level generative models. PartCrafter jointly generates multiple components with compositional 3D latents (Lin et al., 2025); OmniPart combines structure planning with spatially conditioned part synthesis (Yang et al., 2025); and AutoPartGen generates object parts autoregressively (Chen et al., 2025b). From 3D inputs, Hunyuan3D-Part integrates native 3D part segmentation with controllable reconstruction (Ma et al., 2025; Yan et al., 2025). More recent work further improves part representation and structural modeling: FullPart preserves fine geometry through full-resolution part generation (Ding et al., 2026), hierarchical hyperbolic modeling captures part relations (Yang et al., 2026a), and UniPart jointly models geometry and part structure in a unified geometry–segmentation latent space (He et al., 2026). Despite substantial progress, existing methods largely overlook how decomposition affects generation difficulty and quality. We address this gap with PDA, which evaluates part-wise generation difficulty using surface complexity and occlusion rate, and MCS, which iteratively refines the initial partition under structural constraints.

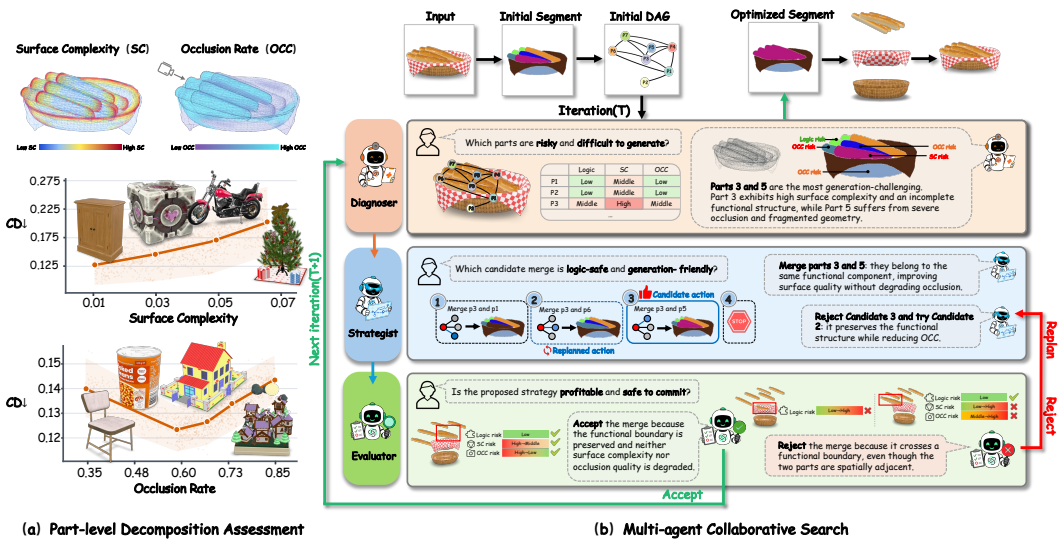


Figure 2: **Overview of FreePart.** A generation-difficulty-aware VLM supports three complementary agents that collaboratively optimize an initial part decomposition through DAG-based search. Candidate refinements are diagnosed, proposed, and evaluated in an iterative accept-or-rollback loop, progressively producing a decomposition better suited to part-aware 3D generation.

3 METHOD

To identify generation-friendly decompositions, we first formulate the task as an optimization problem over candidate part decompositions, with the goal of improving downstream 3D generation quality. To guide this optimization, we introduce Part-level Decomposition Assessment (PDA), which characterizes part-level generation difficulty through surface complexity (SC) and occlusion rate (OCC). Building on PDA, we propose Multi-agent Collaborative Search (MCS), where we fine-tune a shared VLM to recognize geometric and spatial risks and guide decomposition decisions. Three agents collaborate over a decomposition DAG to progressively refine the initial partition.

3.1 PROBLEM FORMULATION

Given an input observation I, holistic 3D generation predicts a complete 3D object using a generation model G:

$$\Phi = G(I), \tag{1}$$

where Φ denotes the generated object. Here, I depends on the downstream generation pipeline.

Part-based generation instead starts from an initial part configuration $\mathcal{P}_0 = \{P_i^0\}_{i=1}^{K_0}$. For a given configuration $\mathcal{P} = \{P_i\}_{i=1}^{K}$, each part is generated independently from its corresponding observation I_i and assembled into a complete object:

$$\phi(\mathcal{P}) = \Gamma\left(\{G(I_i)\}_{i=1}^{K}\right), \tag{2}$$

where Γ denotes the assembly operator. With G fixed, different semantically valid part configurations may lead to substantially different generation quality.

We therefore seek a generation-friendly configuration:

$$\mathcal{P}^\star = \arg\max_{\mathcal{P}\in\Omega(\mathcal{P}_0)} Q\left(\phi(\mathcal{P})\right), \tag{3}$$

where $\Omega(\mathcal{P}_0)$ denotes the set of configurations reachable from the initial decomposition $\mathcal{P}_0$ through successive adjacent-part merges, and $Q(\cdot)$ evaluates the quality of the assembled 3D result.

3.2 PART-LEVEL DECOMPOSITION ASSESSMENT

Existing methods primarily determine part decompositions based on semantic and structural plausibility, yet equally plausible decompositions can differ markedly in downstream generation quality. To search for generation-friendly part decompositions, we model two key factors—Surface Complexity (SC) and Occlusion Rate (OCC).

Surface Complexity. To explore how geometric characteristics may influence 3D generation difficulty, we first investigate Surface Complexity (SC). Intuitively, parts with flat or smoothly varying surfaces may be easier to reconstruct, whereas sharp creases, dense curvature variations, and irregular structures may pose greater challenges. This observation motivates us to examine the relationship between surface complexity and generation quality.

We measure part-level surface complexity using local normal variation between adjacent mesh faces. For a part $\mathcal{P}_p$, let $\mathcal{E}_p^{\text{int}}$ denote its internal adjacent-face pairs, l_{ij} the shared-edge length, and $\mathbf{n}_i, \mathbf{n}_j$ the corresponding unit normals. Its surface complexity is defined as

$$\theta_{ij} = \operatorname{atan2}\left(\|\mathbf{n}_i \times \mathbf{n}_j\|_2, \mathbf{n}_i^\top \mathbf{n}_j\right),$$
$$C_{\text{part}}(\mathcal{P}_p) = \frac{\sum_{(i,j)\in\mathcal{E}_p^{\text{int}}} l_{ij}\theta_{ij}}{\pi \sum_{(i,j)\in\mathcal{E}_p^{\text{int}}} l_{ij}}. \tag{4}$$

A larger value indicates stronger surface variation and potentially higher generation difficulty.

Occlusion Rate. Beyond surface complexity, generation difficulty is also affected by the amount of geometric information available from the input view. We therefore define Occlusion Rate (OCC) based on the proportion of 3D surface area that is unobservable from this view.

We further characterize the amount of geometric information available from the input view using 3D surface occlusion. Let $A(\mathcal{P}_i)$ denote the surface area of part $\mathcal{P}_i$ and $A_{\text{vis}}(\mathcal{P}_i \mid \mathcal{M}, v)$ its visible area from view v when evaluated within the complete mesh $\mathcal{M}$. The part-level occlusion rate is

$$\text{Occ}_i(v) = 1 - \frac{A_{\text{vis}}(\mathcal{P}_i \mid \mathcal{M}, v)}{A(\mathcal{P}_i)}. \tag{5}$$

Unlike a 2D projected-area ratio, this metric measures the proportion of unobservable 3D surface. Evaluating visibility on the complete mesh captures both self-occlusion and inter-part occlusion. We provide the global definitions of SC and OCC, along with further details, in Appendix A.

Analysis of Generation-Difficulty Assessment. We analyze how surface complexity and occlusion relate to downstream 3D generation quality. As shown in Fig. 2 (a), the mean Chamfer Distance (CD) generally increases with surface complexity, indicating that geometrically complex parts tend to be more difficult to reconstruct. Occlusion shows a different, non-monotonic relationship with reconstruction quality: generation error is lower within an intermediate occlusion range and increases when the occlusion rate deviates from this range. These observations suggest that SC and OCC provide complementary indicators of part-level generation difficulty.

3.3 MULTI-AGENT COLLABORATIVE SEARCH

Generation-Aware VLM Optimization. Guided by PDA, we develop Multi-agent Collaborative Search (MCS) to progressively refine the initial part decomposition before downstream generation. MCS employs three complementary agents for risk diagnosis, candidate selection, and outcome evaluation. To provide these roles with a consistent understanding of part-level generation difficulty, we optimize a shared vision-language model (VLM) in two stages.

We first construct generation-difficulty supervision from the SC and OCC measures introduced in PDA. For each training part P_i, its surface-complexity risk and occlusion risk are obtained by discretizing the corresponding PDA measurements:

$$\rho_i^{\text{SC}} = Q_{\text{SC}}(C_{\text{part}}(P_i)), \qquad \rho_i^{\text{OCC}} = Q_{\text{OCC}}(|\text{Occ}_i(v) - \tau_{\text{occ}}|), \tag{6}$$

where Q_{SC} and Q_{OCC} map the continuous measurements into $\{\mathrm{low}, \mathrm{medium}, \mathrm{high}\}$ risk levels, and τ_{occ} denotes the empirically favorable occlusion level. We further define the overall generation risk as

$$\rho_i^G = \max_{\prec} \left\{ \rho_i^{\mathrm{SC}}, \rho_i^{\mathrm{OCC}} \right\}, \tag{7}$$

where $\mathrm{low} \prec \mathrm{medium} \prec \mathrm{high}$.

These three labels form the supervision target

$$\mathbf{y}_i = \left(\rho_i^{\mathrm{SC}}, \rho_i^{\mathrm{OCC}}, \rho_i^G \right). \tag{8}$$

Using $\mathbf{y}_i$ as supervision, supervised fine-tuning (SFT) (Ouyang et al., 2022) trains the shared VLM to predict part-level generation risks from a complete-object rendering I and the corresponding target-part mask M_i. The 3D-derived SC and OCC measurements are used only for training supervision and are not required at inference.

Starting from this risk-aware VLM, group-based policy optimization (GDPO) (Liu et al., 2026a) further optimizes the shared VLM with role-specific prompts and rewards, encouraging each role to preserve part logic while reducing SC- and OCC-related generation risks. Detailed reward definitions and optimization objectives are provided in Appendix B.

Complementary Agent Roles. At each iteration, the agents operate sequentially under a shared controller. The Diagnoser assesses the SC, OCC, and logical completeness of the current parts, identifying high-risk or fragmented regions without directly deciding how to merge them. Conditioned on this diagnosis, the Strategist explores the remaining candidate transitions and proposes the next merge, or predicts stop. It considers logical validity and the expected changes in both risks, favoring safe candidates on the SC and OCC. The Evaluator independently verifies the selected action by comparing the decomposition before and after the virtual merge, without accessing the reasoning of the preceding agents. It accepts a merge only if part-level logic is preserved and neither predicted SC nor OCC risk increases; otherwise, the proposal is rejected. The responsibilities and collaboration of the three agents are illustrated in Fig. 2(b). Formally, the roles interact as

$$\begin{aligned} D_t &= \pi_{\mathrm{diag}}(I, \mathcal{S}_t), \\ a_t &= \pi_{\mathrm{str}}(I, \mathcal{S}_t, D_t, \mathcal{A}_t), \\ v_t &= \pi_{\mathrm{eval}}\left(I, \mathcal{S}_t, \widetilde{\mathcal{S}}_{t+1}, a_t\right), \end{aligned} \tag{9}$$

where D_t is the diagnosis, $\mathcal{A}_t$ is the available candidate set, $\widetilde{\mathcal{S}}_{t+1}$ is the virtually merged state, and $v_t \in \{\texttt{accept}, \texttt{rollback}\}$. The Evaluator is invoked only for a proposed merge. During inference, all risk judgments are made from the available observations by the shared VLM, without access to ground-truth 3D SC or OCC values.

DAG-Guided Collaborative Search. As illustrated in Fig. 3, we organize this collaborative process as a directed acyclic graph (DAG) search rather than predicting the final decomposition in one step. Each node represents a decomposition state $\mathcal{S}_t = \{\mathcal{P}_1^t, \ldots, \mathcal{P}_{K_t}^t\}$, and each directed edge is a pairwise merge a_{ij}^t leading to

$$T(\mathcal{S}_t, a_{ij}^t) = \left(\mathcal{S}_t \setminus \{\mathcal{P}_i^t, \mathcal{P}_j^t\}\right) \cup \{\mathcal{P}_i^t \cup \mathcal{P}_j^t\}. \tag{10}$$

Each accepted merge reduces the number of parts by one, making the graph acyclic. Rather than constructing the full DAG, the controller instantiates only the outgoing edges of the current state. Adjacency first removes merges between spatially unrelated parts; the Diagnoser then prunes branches judged irrelevant or unsafe according to its risk and logic assessment. Previously rejected edges are masked at the current node:

$$\begin{aligned} \mathcal{A}_t^{\mathrm{adj}} &= \left\{ a_{ij}^t \mid \mathrm{Adj}(\mathcal{P}_i^t, \mathcal{P}_j^t) = 1 \right\}, \\ \mathcal{A}_t &= \mathrm{Prune}\left(\mathcal{A}_t^{\mathrm{adj}}; D_t\right) \setminus \mathcal{B}_t, \end{aligned} \tag{11}$$

where B_t is the rejected-action set, and $\mathrm{Prune}(\cdot)$ filters out unsuitable candidate merges according to D_t. The Strategist selects a candidate edge from $\mathcal{A}_t$, proposing a one-step extension of the current

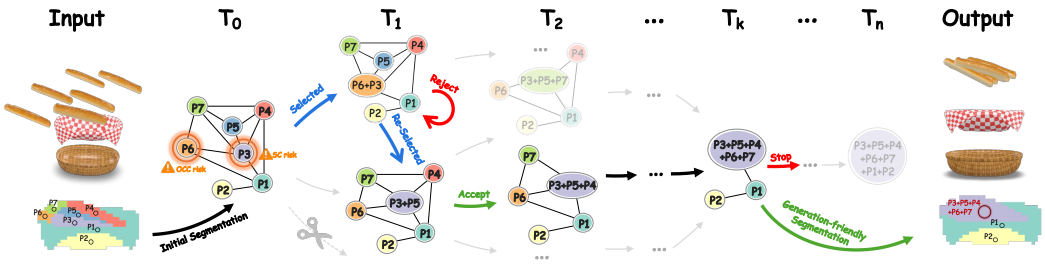


Figure 3: Illustration of the DAG-based decomposition search. Starting from the initial segmentation (decomposition), SC/OCC risks guide the selection of adjacent-part merge candidates. Unreliable transitions are rejected and reselected, infeasible branches are pruned, and accepted transitions iteratively update the decomposition until the stopping criterion is reached, yielding the optimized segmentation for downstream 3D generation.

search path. This edge is previewed without changing $\mathcal{S}_t$; the Evaluator then determines whether the search advances or returns to Strategist:

$$(\mathcal{S}_{t+1}, \mathcal{B}_{t+1}) = \begin{cases} (\widetilde{\mathcal{S}}_{t+1}, \varnothing), & v_t = \texttt{accept}, \\ (\mathcal{S}_t, \mathcal{B}_t \cup \{a_t\}), & v_t = \texttt{rollback}. \end{cases} \tag{12}$$

An accepted merge advances to the successor node, where adjacency is updated and a new diagnosis begins. A rejected merge leaves the current state unchanged; the controller masks that edge and asks Strategist to propose an alternative using the existing diagnosis. Repeating this diagnosis–strategy–evaluation cycle builds a path toward a decomposition better suited to downstream generation. Search terminates when Strategist predicts `stop`, no admissible edge remains, or the iteration limit is reached. Modality-specific adjacency tests and search details are provided in Appendix C.

4 Experiments

4.1 Settings

Implementation Details. We use Qwen2.5-VL-72B-Instruct as the shared backbone for all three agents. We first perform generation-difficulty-aware SFT on 50K part-level samples using 4-bit QLoRA, and subsequently optimize the resulting model with GDPO under the logic, SC, and OCC rewards described in Sec. 3. All experiments are conducted on eight NVIDIA A800 GPUs. Full training and inference configurations are provided in the appendix.

Datasets. We evaluate FreePart on three benchmarks covering diverse inputs and generation scales: PartObjaverse-Tiny (Yang et al., 2024) and PartNeXt (Wang et al., 2025) for image-to-3D and mesh-to-3D generation, and 3D-Front-MIDI (Huang et al., 2025a) for image-to-3D scene generation.

Evaluation Metrics. Following OmniPart, we evaluate decomposition accuracy using Bbox IoU and generation quality using Chamfer Distance (CD) and F-score at thresholds of 0.1 and 0.05. CD and F-score are reported at both the part and overall-object levels.

Baselines. We integrate FreePart into three representative part-aware generation frameworks: OmniPart (Yang et al., 2025) for image-to-3D, Hunyuan3D-Part (Ma et al., 2025; Yan et al., 2025) for mesh-to-3D, and MIDI-3D (Huang et al., 2025a) for image-to-3D scenes. The original frameworks are retained as baselines to evaluate FreePart across different backbones and input modalities.

4.2 Main Results

Table 1 demonstrates the consistent effectiveness of FreePart across diverse part-aware 3D generation settings. When integrated with OmniPart, Hunyuan3D-Part, and MIDI-3D, FreePart improves both decomposition quality and 3D reconstruction performance on all evaluated datasets. Its consistent gains across image-, mesh-, and scene-conditioned generation demonstrate that FreePart serves as a general decomposition refinement framework rather than being specialized to a particular backbone or modality. The qualitative comparisons in Fig. 4 further support this finding: while the

Table 1: Quantitative comparison of FreePart (FP) across different part-aware 3D generation tasks and datasets. We evaluate decomposition accuracy and generation quality at both the part and overall levels.

Task	Method	Dataset	Segmentation	Part-level			Overall-level		
			Bbox IoU ↑	CD ↓	F1-0.1 ↑	F1-0.05 ↑	CD ↓	F1-0.1 ↑	F1-0.05 ↑
Image-to-3D	OmniPart	PartObjaverse-Tiny	23.31	27.37	48.44	26.94	14.80	77.30	52.21
	OmniPart+FP		**25.07**	**24.69**	**53.72**	**31.06**	**14.13**	**79.85**	**55.05**
	OmniPart	PartNeXt	30.55	32.70	42.30	25.29	13.23	84.14	67.88
	OmniPart+FP		**38.13**	**29.87**	**47.88**	**30.00**	**12.79**	**85.07**	**69.87**
Mesh-to-3D	Hunyuan3D-Part	PartObjaverse-Tiny	50.70	22.69	58.60	37.42	7.91	93.80	77.56
	Hunyuan3D-Part+FP		**51.67**	**21.96**	**60.30**	**38.76**	**6.87**	**96.64**	**82.95**
	Hunyuan3D-Part	PartNeXt	9.10	26.65	53.53	36.06	6.00	95.56	85.15
	Hunyuan3D-Part+FP		**11.13**	**25.82**	**55.35**	**38.20**	**5.33**	**97.05**	**88.60**
Image-to-3D Scene	MIDI-3D	3D-Front-MIDI	23.77	23.88	54.84	30.69	23.68	57.25	34.46
	MIDI-3D+FP		**26.08**	**21.62**	**59.65**	**34.44**	**22.23**	**60.43**	**37.21**

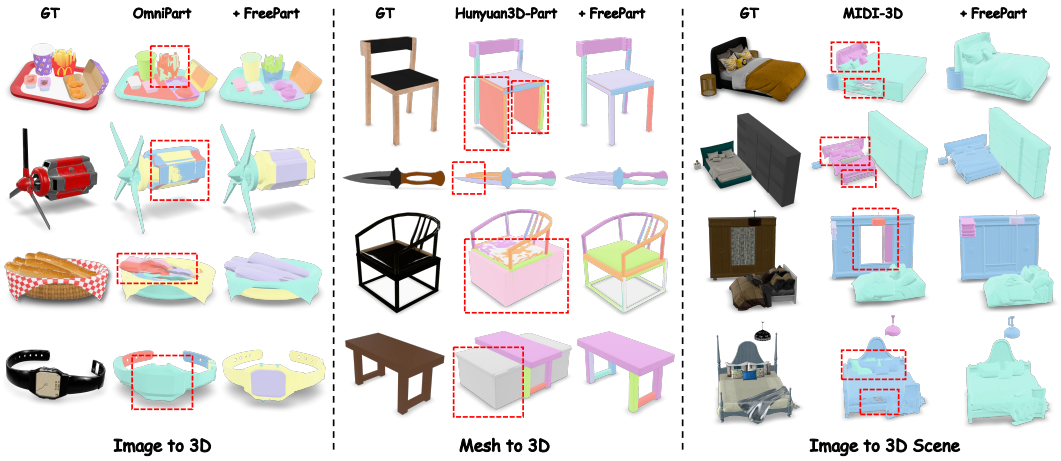


Figure 4: Qualitative comparisons across image-to-3D, mesh-to-3D, and image-to-3D scene generation. Red boxes highlight incomplete or structurally inconsistent regions in the baseline results, while FreePart produces more coherent part geometry under the same downstream generators.

baselines exhibit missing or structurally inconsistent geometry in the highlighted regions, FreePart produces more complete and coherent parts under the same downstream generators. Together, these results demonstrate that refining the input decomposition reliably improves downstream 3D generation without modifying the generation backbone.

User Study. We conduct a pairwise user study with 18 participants and 540 comparisons to evaluate the perceptual quality of our results against the corresponding baselines. As shown in Table 2, our method is preferred in 62.2% of all comparisons, compared with 25.9% for the baselines, and consistently leads across all three generation settings. The preferences in Mesh-to-3D and Image-to-3D Scene further demonstrate the perceptual benefits and cross-task generalizability of decomposition optimization.

Table 2: User study results

Setting	Ours	Baseline	Tie
Image-to-3D	**44.4%**	34.3%	21.2%
Mesh-to-3D	**76.5%**	19.8%	3.7%
Image-to-3D Scene	**68.9%**	22.2%	8.9%
Overall	**62.2%**	25.9%	11.9%

4.3 Ablation Study

We conduct ablations on three components: risk indicators, shared-VLM fine-tuning, and the multi-agent pipeline. All ablation experiments are conducted on a fixed subset of PartObjaverse-Tiny with OmniPart under the same evaluation setting. Within each ablation group, all variants use identical

test samples and evaluation schemes, with other settings held fixed. The full model achieves the best overall performance, demonstrating that these components play complementary roles in PDA-guided MCS and jointly improve downstream generation quality.

Effect of SC and OCC. To validate whether SC and OCC can effectively guide decomposition optimization toward better 3D generation, we conduct a controlled ablation study on top of a strong baseline. We vary only the indicator signals available to the decomposition pipeline, thereby isolating their effects without introducing differences from model capacity or training. As shown in Table 3, introducing either SC or OCC consistently improves all metrics: SC identifies geometrically complex parts, while OCC captures insufficient visual evidence from the input view. Their combination achieves the best overall performance, demonstrating that the two indicators provide complementary guidance for refining part decompositions.

Table 3: Effect of SC and OCC

Indicator		Overall-level		
SC	OCC	CD ↓	F1-0.1 ↑	F1-0.05 ↑
×	×	14.08	80.99	55.16
✓	×	13.71	81.68	57.01
×	✓	13.71	81.72	56.89
✓	✓	**13.48**	**82.26**	**57.86**

Effect of VLM Fine-tuning. To examine whether progressive VLM training improves both risk perception and decomposition decision-making, we conduct an ablation study of the shared VLM. As shown in Table 4, SFT improves upon the base VLM by introducing explicit supervision for part-level structural understanding and risk perception. GDPO further enhances performance by aligning the acquired perception with multi-objective rewards, enabling more reliable decisions among candidate decomposition paths. The consistent progression across the three variants demonstrates that SFT establishes task-specific perception, while GDPO converts this capability into effective search guidance.

Table 4: Effect of VLM Fine-tuning

VLM		Overall-level		
SFT	GDPO	CD ↓	F1-0.1 ↑	F1-0.05 ↑
×	×	13.89	81.24	56.39
✓	×	13.68	82.04	56.96
✓	✓	**13.48**	**82.26**	**57.86**

Effect of the Multi-Agent Pipeline. To determine whether the performance gains arise from specialized agent collaboration, we conduct a controlled ablation study of the multi-agent pipeline. As shown in Table 5, removing any of the three capabilities degrades the final performance. *Diagnoser.* Removing risk warnings deprives the search of targeted guidance. *Strategist.* Replacing adaptive proposals with fixed rules prevents merge decisions from adapting to the current decomposition. *Evaluator.* Disabling rollback weakens error correction, allowing erroneous merges to persist across iterations. The full system achieves the strongest results by forming a closed-loop search process, which confirms that the three agents are complementary rather than interchangeable, jointly enabling robust and adaptive DAG-based decomposition.

Table 5: Effect of the Multi-Agent Pipeline

Agent Variant	Overall-level		
	CD ↓	F1-0.1 ↑	F1-0.05 ↑
Diagnoser w/o warning	14.33	79.96	54.02
Strategist w/ fixed rules	14.18	80.41	54.74
Evaluator w/o rollback	13.83	81.50	55.54
Full system	**13.48**	**82.26**	**57.86**

5 Conclusion and Limitation

In this work, we revisit part decomposition as an optimizable component of part-level 3D generation rather than fixed preprocessing. We present FreePart, a plug-and-play framework that optimizes part decomposition for downstream 3D generation. FreePart combines Part-level Decomposition Assessment (PDA) with Multi-agent Collaborative Search (MCS) to evaluate candidate decompositions and progressively refine them. Experiments across multiple pipelines and benchmarks demonstrate consistent improvements in part-level coherence and overall-object completeness.

FreePart remains dependent on the initial decomposition and may not fully recover missing parts or severe boundary errors. Moreover, PDA considers only selected geometric and spatial factors, while iterative search introduces additional inference costs. Future work will explore joint segmentation and decomposition optimization with richer structural cues and more efficient search.

AI USE STATEMENT

In this work, we used generative AI tools primarily for language polishing and improving the readability and presentation of the manuscript. Generative AI tools provided limited assistance with manuscript organization, formatting, and visual presentation. All AI-assisted materials were manually reviewed and verified by the authors, who take full responsibility for the final content, claims, and results of this work.

REPRODUCIBILITY STATEMENT

To facilitate verification of our method, we provide anonymized supplementary code covering the core FreePart pipeline, including geometry-aware supervised fine-tuning, three-role GDPO optimization, multi-agent DAG-based decomposition search, and integration and evaluation with OmniPart. The supplementary package contains data-processing utilities, training and inference entry points, experiment configurations, fixed evaluation manifests, and tests for the principal algorithmic components. Required sources, expected directory structures, preprocessing procedures, and available artifact checksums are documented in the supplementary material. These resources support independent inspection of FreePart and reproduction of the released OmniPart-based pipeline under the documented experimental environment.

REFERENCES

Minghao Chen, Roman Shapovalov, Iro Laina, Tom Monnier, Jianyuan Wang, David Novotny, and Andrea Vedaldi. PartGen: Part-level 3D generation and reconstruction with multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5881–5892, 2025a.

Minghao Chen, Jianyuan Wang, Roman Shapovalov, Tom Monnier, Hyunyoung Jung, Dilin Wang, Rakesh Ranjan, Iro Laina, and Andrea Vedaldi. AutoPartGen: Autoregressive 3D part generation and discovery. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J. Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jia-Wei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3D: 3Dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026.

Lihe Ding, Shaocong Dong, Yaokun Li, Chenjian Gao, Xiao Chen, Rui Han, Yihao Kuang, Hong Zhang, Bo Huang, Zhanpeng Huang, Zibin Wang, Dan Xu, and Tianfan Xue. FullPart: Generating each 3D part at full resolution. In *International Conference on Learning Representations*, 2026.

Xufan He, Yushuang Wu, Xiaoyang Guo, Chongjie Ye, Jiaqing Zhou, Tianlei Hu, Xiaoguang Han, and Dong Du. UniPart: Part-level 3D generation with unified 3D geom-seg latents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34227–34236, 2026.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *International Conference on Learning Representations*, 2024.

Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. MIDI: Multi-instance diffusion for single image to 3D scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23646–23657, 2025a.

Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M. Rehg, and Varun Jampani. SPAR3D: Stable point-aware reconstruction of 3D objects from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16860–16870, 2025b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 300–309, 2023.

Yuchen Lin, Chenguo Lin, Panwang Pan, Honglei Yan, Yiqiang Feng, Yadong Mu, and Katerina Fragkiadaki. PartCrafter: Structured 3D mesh generation via compositional latent diffusion transformers. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Anran Liu, Cheng Lin, Yuan Liu, Xiaoxiao Long, Zhiyang Dou, Hao-Xiang Guo, Ping Luo, and Wenping Wang. Part123: Part-aware 3D reconstruction from a single-view image. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. doi: 10.1145/3641519.3657482.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. In *International Conference on Machine Learning*, 2026a.

Xinyue Liu, Jin Liu, Hongbo Wang, Ran He, and Huaibo Huang. Think-then-generate: Structural chain-of-thought reasoning for consistent 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34270–34280, 2026b.

Changfeng Ma, Yang Li, Xinhao Yan, Jiachen Xu, Yunhan Yang, Chunshi Wang, Zibo Zhao, Yanwen Guo, Zhuo Chen, and Chunchao Guo. P3-SAM: Native 3D part segmentation. arXiv preprint arXiv:2509.06784, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations*, 2023.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *International Conference on Learning Representations*, 2024.

Penghao Wang, Yiyang He, Xin Lv, Yukai Zhou, Lan Xu, Jingyi Yu, and Jiayuan Gu. PartNeXt: A next-generation dataset for fine-grained and hierarchical 3D part understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2030. Datasets and Benchmarks Track.

Xiwen Wang, Shichao Zhang, Ruowei Wang, Mao Li, Chenyu Zhou, Ji-Zhe Zhou, Qijun Zhao, and Hailun Zhang. Dehallu3D: Hallucination-mitigated 3D generation from a single image via cyclic view consistency refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19993–20002, 2026.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, and Lu Sheng. Ouroboros3D: Image-to-3D generation via 3D-aware recursive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21631–21641, 2025.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21469–21480, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14419–14429, 2026.

Xinhao Yan, Jiachen Xu, Yang Li, Changfeng Ma, Yunhan Yang, Chunshi Wang, Zibo Zhao, Zeqiang Lai, Yunfei Zhao, Zhuo Chen, and Chunchao Guo. X-Part: High fidelity and structure coherent shape decomposition. arXiv preprint arXiv:2509.08643, 2025.

Qitong Yang, Mingtao Feng, Zijie Wu, Huixin Zhu, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Learning hierarchical hyperbolic mixture model for part-aware 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2026a.

Yunhan Yang, Yukun Huang, Yuan-Chen Guo, Liangjun Lu, Xiaoyang Wu, Edmund Y. Lam, Yan-Pei Cao, and Xihui Liu. SAMPart3D: Segment any part in 3D objects. *arXiv preprint arXiv:2411.07184*, 2024.

Yunhan Yang, Yufan Zhou, Yuan-Chen Guo, Zi-Xin Zou, Yukun Huang, Ying-Tian Liu, Hao Xu, Ding Liang, Yan-Pei Cao, and Xihui Liu. OmniPart: Part-aware 3D generation with semantic decoupling and structural cohesion. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025. doi: 10.1145/3757377.3763872.

Yunhan Yang, Yuan-Chen Guo, Yukun Huang, Zi-Xin Zou, Zhipeng Yu, Yangguang Li, Yan-Pei Cao, and Xihui Liu. HoloPart: Generative 3D part amodal segmentation. In *International Conference on Learning Representations*, 2026b.

A Generation-Difficulty Assessment

Surface Complexity Given a triangular mesh $\mathcal{M} = (\mathcal{V}, \mathcal{F})$, we define the set of adjacent face pairs as

$$\mathcal{E}_{\text{adj}} = \{(i,j) \mid f_i, f_j \in \mathcal{F},\ f_i \cap f_j = e_{ij}\}, \tag{13}$$

where e_{ij} is the shared edge with length l_{ij}. For any valid $\mathcal{E} \subseteq \mathcal{E}_{\text{adj}}$, the unified surface complexity operator is

$$\mathcal{C}(\mathcal{E}) = \frac{\sum_{(i,j)\in\mathcal{E}} l_{ij}\theta_{ij}}{\pi \sum_{(i,j)\in\mathcal{E}} l_{ij}}. \tag{14}$$

Edge-length weighting emphasizes variations extending over larger surface regions and reduces the influence of short mesh edges. The normalization bounds $\mathcal{C}(\mathcal{E})$ within $[0,1]$.

Let s_i denote the part label of face f_i. The global and part-level measures are obtained using different aggregation domains:

$$\begin{aligned} C_{\text{global}}(\mathcal{M}) &= \mathcal{C}(\mathcal{E}_{\text{adj}}), \\ \mathcal{E}_p^{\text{int}} = \{(i,j) \in \mathcal{E}_{\text{adj}} \mid s_i = s_j = p\}, &\qquad C_{\text{part}}(\mathcal{P}_p) = \mathcal{C}(\mathcal{E}_p^{\text{int}}). \end{aligned} \tag{15}$$

Face pairs with $s_i \neq s_j$ are excluded from the part-level measure to avoid conflating within-part geometry with partition-induced boundary effects. These cross-part interfaces are instead retained for constructing part adjacency and candidate merge actions. After a candidate merge, the internal adjacency set is updated according to the new part labels.

Surface Occlusion For a mesh $\mathcal{M}$ with surface $\mathcal{S}(\mathcal{M})$, its total surface area and visible area under view v are

$$\begin{aligned} A(\mathcal{M}) &= \int_{\mathcal{S}(\mathcal{M})} dA, \\ A_{\text{vis}}(\mathcal{M}, v) &= \int_{\mathcal{S}(\mathcal{M})} \mathbf{1}\left[\mathbf{x} \in \mathcal{V}(\mathcal{M}, v)\right] dA, \end{aligned} \tag{16}$$

where $\mathcal{V}(\mathcal{M}, v)$ denotes the visible surface. The object-level occlusion rate is

$$\mathrm{Occ}_{\mathrm{obj}}(\mathcal{M}, v) = 1 - \frac{A_{\mathrm{vis}}(\mathcal{M}, v)}{A(\mathcal{M})}. \tag{17}$$

For a part $\mathcal{P}_i$, visibility is evaluated against the complete mesh:

$$A_{\mathrm{vis}}(\mathcal{P}_i \mid \mathcal{M}, v) = \int_{\mathcal{S}(\mathcal{P}_i)} \mathbf{1}\left[\mathbf{x} \in \mathcal{V}(\mathcal{M}, v)\right] dA. \tag{18}$$

This formulation accounts for both surfaces hidden by the part itself and surfaces occluded by other parts.

Since continuous visibility is difficult to compute analytically, we estimate the visible area through area-weighted surface sampling. Samples are drawn according to triangle areas and evaluated under the input view using depth-consistency tests, and face-index tests. The visible area is obtained by summing the area weights of visible samples.

B VLM Optimization

SFT Targets and Objective. The categorical targets in Eqs. equation 19 and equation 20 are constructed from the part-level SC and OCC measures defined in Sec. 3. We discretize surface complexity and deviation from the empirically favorable occlusion rate using Q_{SC} and Q_{OCC}:

$$\rho_i^{\mathrm{SC}} = Q_{\mathrm{SC}}(C_{\mathrm{part}}(\mathcal{P}_i)), \qquad \rho_i^{\mathrm{OCC}} = Q_{\mathrm{OCC}}(|\mathrm{Occ}_i(v) - \tau_{\mathrm{occ}}|), \tag{19}$$

$$\rho_i^{\mathrm{G}} = \max_{\prec} \left\{\rho_i^{\mathrm{SC}}, \rho_i^{\mathrm{OCC}}\right\}, \tag{20}$$

where $\mathrm{low} \prec \mathrm{medium} \prec \mathrm{high}$ and $\tau_{\mathrm{occ}} = 0.6$. Thus, a lower OCC risk means closer proximity to the empirically favorable occlusion range, rather than a smaller raw occlusion rate. The reference value and discretization boundaries are determined on the training split and kept fixed for validation and testing.

Given a complete-object rendering I and its target-part mask M_i, the structured response $\mathbf{y}_i$ is learned through the token-level SFT objective

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_i \sum_{s=1}^{T_i} \log p_\theta(y_{i,s} \mid I, M_i, y_{i,<s}). \tag{21}$$

The mesh-derived SC and OCC values provide training supervision; their ground-truth values are not supplied to the agents at inference time.

Role-Specific GDPO Rewards. The shared VLM is subsequently optimized with prompts for Diagnoser, Strategist, and Evaluator. Each sampled response receives a three-dimensional reward

$$\mathbf{r}_k^{(j)} = \left(r_{k,\mathrm{logic}}^{(j)}, r_{k,\mathrm{SC}}^{(j)}, r_{k,\mathrm{OCC}}^{(j)}\right), \qquad j \in \{\mathrm{D}, \mathrm{S}, \mathrm{E}\}. \tag{22}$$

For Diagnosis, the rewards assess part-level risk and logic predictions; for Strategy, they assess the proposed action and its predicted effects; and for Evaluation, they assess the accept-or-rollback decision and the predicted changes.

For a candidate merge a, let $L(a) \in \{0, 1\}$ indicate whether it preserves functional part boundaries. During training, this label is derived from the PartNeXt functional-instance annotations. Let $R_d^{\mathrm{before}}(a)$ and $R_d^{\mathrm{after}}(a)$ denote the ordered risk levels before and after the tentative merge for $d \in \{\mathrm{SC}, \mathrm{OCC}\}$. We use their difference to indicate the direction of change:

$$\Delta_d(a) = R_d^{\mathrm{before}}(a) - R_d^{\mathrm{after}}(a). \tag{23}$$

For the candidate set $\mathcal{A}$ in a training example, safe and beneficial actions satisfy

$$\mathcal{A}_{\mathrm{safe}} = \{a \in \mathcal{A} \mid L(a) = 1,\ \Delta_{\mathrm{SC}}(a) \geq 0,\ \Delta_{\mathrm{OCC}}(a) \geq 0,\ \Delta_{\mathrm{SC}}(a) + \Delta_{\mathrm{OCC}}(a) > 0\}. \tag{24}$$

Strategy targets are the non-dominated actions on the SC–OCC Pareto front of $\mathcal{A}_{\mathrm{safe}}$; the target is `stop` when this set is empty. At inference time, these judgments are made from predicted risks and logic rather than training annotations.

C DAG-BASED DECOMPOSITION SEARCH

Decomposition DAG and Lazy Construction. We represent the reachable decompositions as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where a node is a partition of the input object into mutually exclusive parts and an edge is an adjacent-part merge, as defined in Eq. equation 10. Each accepted edge decreases the part count by one; therefore, no directed cycle is possible. Distinct merge orders may reach the same partition, which is represented as one conceptual DAG node. Because enumerating all reachable partitions is computationally expensive, the controller instantiates only the outgoing actions of the current state and updates them after an accepted merge.

Modality-Specific Adjacency. For image-based inputs, two parts are adjacent when their mask boundaries contact or lie within a small tolerance. For mesh-based inputs, we use face adjacency along a shared mesh edge:

$$\mathrm{Adj}_{\mathrm{2D}}(\mathcal{P}_i, \mathcal{P}_j) = \mathbb{I}[d(\partial M_i, \partial M_j) \le \epsilon_{\mathrm{2D}}]\,, \tag{25}$$

$$\mathrm{Adj}_{\mathrm{3D}}(\mathcal{P}_i, \mathcal{P}_j) = \mathbb{I}[\exists f_p \in \mathcal{F}_i,\ f_q \in \mathcal{F}_j : f_p \cap f_q = e_{pq}]\,. \tag{26}$$

Here M_i and M_j are part masks, and ϵ_{2D} is the image-space boundary tolerance. These tests instantiate the adjacency-valid edges $\mathcal{A}_t^{\mathrm{adj}}$ in Eq. equation 11; no non-adjacent merge is proposed.

Diagnosis-Conditioned Candidate Management. At the current DAG node, the Diagnoser produces D_t, including part-level SC, OCC, and logical-risk assessments. As specified in Eq. equation 11, diagnosis-conditioned pruning acts on $\mathcal{A}_t^{\mathrm{adj}}$, and the controller then masks the rejected-action set $\mathcal{B}_t$. The Diagnoser narrows the local candidate set without selecting a merge; the Strategist proposes the next edge from the remaining actions. The mask $\mathcal{B}_t$ applies only at the current node. It prevents repeated evaluation of a rejected transition there but does not exclude an action at a different decomposition state. All risk signals used for inference are VLM predictions.

Tentative Transition and Rollback. A Strategy proposal is previewed before it changes the search state:

$$\widetilde{\mathcal{S}}_{t+1} = T(\mathcal{S}_t, a_t). \tag{27}$$

The Evaluator compares $\mathcal{S}_t$ with $\widetilde{\mathcal{S}}_{t+1}$ and returns an accept-or-rollback verdict. The controller applies the state update in Eq. equation 12. Acceptance commits the successor state, clears the rejected-action mask, recomputes adjacency, and initiates a new diagnosis. Rollback keeps $\mathcal{S}_t$ and the existing diagnosis D_t, adds a_t to $\mathcal{B}_t$, and asks the Strategist to select another outgoing edge. Thus, a rejected preview is not a reverse traversal of the DAG.

Stopping Criteria and Finite Termination. Search stops if Strategist predicts `stop`, no admissible edge remains, or the iteration limit is reached. At a state with K_t parts, there are at most $\binom{K_t}{2}$ pairwise merges before adjacency and diagnosis-based pruning. Every accepted action decreases the part count, and every rejected action removes one edge from the current candidate set. Thus, the search can accept at most $K_0 - 1$ merges and, ignoring pruning, evaluate at most

$$\sum_{k=2}^{K_0} \binom{k}{2} \tag{28}$$

candidate actions along a trajectory. The bound is finite; adjacency and diagnosis-based pruning further reduce the candidates explored in practice.

D IMPLEMENTATION

Shared VLM backbone. We use Qwen2.5-VL-72B-Instruct as the shared backbone for the Diagnoser, Strategist, and Evaluator. The three agents share identical model parameters and are instantiated using role-specific prompts. We freeze the vision encoder and apply LoRA to all linear layers, with rank $r = 8$, scaling factor $\alpha = 16$, and dropout rate 0.05. This introduces 105.3M trainable parameters. All experiments are conducted on eight NVIDIA A800 GPUs.

Generation-difficulty-aware SFT. We construct 50,000 training and 5,000 validation samples, each consisting of a complete-object rendering, a target-part mask, and categorical SC, OCC, and overall generation-risk labels. We fine-tune the backbone for one epoch using 4-bit QLoRA with NF4 double quantization. Computation is performed in `bfloat16`, with FlashAttention and gradient checkpointing enabled.

We use a per-device batch size of 1 and accumulate gradients for 8 steps, resulting in an effective batch size of 64 across eight GPUs. The learning rate is 5×10^{-5} with a cosine schedule, a warm-up ratio of 0.05, and a weight decay of 0.01. The maximum sequence length is 2,048 tokens, and the maximum visual resolution is limited to 262,144 pixels. Training is distributed using DeepSpeed ZeRO-2 with random seed 42. The model is trained for 782 optimization steps, and checkpoint 750 is selected according to validation loss. SFT training takes approximately 5.7 hours.

GDPO training data. The GDPO training set contains 5,537 role-specific examples from 494 PartNeXt objects, including 2,004 Diagnosis, 1,946 Strategy, and 1,587 Evaluation examples. Each Strategy example contains adjacency-constrained candidate merges and their virtual after-merge previews. The held-out validation set contains 300 balanced examples, with 100 examples for each role, collected from 59 objects. The 600 PartNeXt objects reserved for validation and downstream evaluation are excluded from policy training.

Multi-stage GDPO optimization. Starting from the selected SFT checkpoint, we optimize the shared policy using a three-stage GDPO schedule. The first stage is trained for 100 optimization steps with a learning rate of 1×10^{-6}. Starting from its checkpoint at step 100, the second stage uses a learning rate of 5×10^{-7}, from which checkpoint 50 is selected. We then perform a final evaluation-disciplined refinement using a learning rate of 2×10^{-7} and select checkpoint 30 from a 40-step run. The final checkpoint is used unchanged in all downstream generation pipelines.

For each prompt, we sample eight responses using a temperature of 0.9, top-p of 0.9, and top-k of 50. The maximum completion length is 192 tokens. Each response receives three independent rewards for logical correctness, SC, and OCC. The rewards are normalized separately within each response group, aggregated with equal weights, and subsequently whitened across the batch. We use a KL coefficient of 0.04 and a clipping range of 0.2.

Unlike the SFT stage, GDPO training does not use 4-bit quantization. The model is trained in `bfloat16` using DeepSpeed ZeRO-3, FlashAttention, and gradient checkpointing. The per-device batch size is 1 with 8 gradient-accumulation steps. We use a cosine learning-rate schedule without warm-up, a weight decay of 0.1, and random seed 42. The maximum sequence length and visual resolution are 2,048 tokens and 262,144 pixels, respectively.

Effect of LoRA Fine-tuning on Visual Logic Judgement.

To verify whether geometry-risk supervised fine-tuning affects the original part-level visual reasoning ability of the base vision-language model, we construct an additional logic probe to compare the consistency between the original Qwen2.5-VL-72B model and its LoRA-adapted version on candidate part-merging decisions.

Specifically, we sample 100 candidate part-merging cases from held-out objects that are not used during SFT training. Each case contains the complete object image, the part segmentation or candidate merged region, and asks the model to determine whether the proposed merge preserves the visual and functional logic of the object. The model output is constrained to a binary JSON format:

$$\{\texttt{"logic"} : \texttt{"preserved"} \mid \texttt{"violated"}\}. \tag{29}$$

Here, `preserved` indicates that the merge does not break semantic or functional boundaries, while `violated` indicates that the merge may cross independent functional parts or semantic boundaries that should not be merged. We run inference on the same set of cases using both the base Qwen2.5-VL-72B model and the LoRA-adapted model trained with geometry-risk SFT, and then measure their prediction consistency. Note that this experiment does not use human annotations as absolute ground truth. Instead, it uses the base 72B model as a reference to analyze whether LoRA fine-tuning introduces systematic shifts in visual logic judgement.

Table 6 shows the prediction relationship between the base 72B model and the LoRA-adapted 72B model on 100 logic probe cases.

Table 6: Prediction consistency between the base Qwen2.5-VL-72B model and the LoRA-adapted model on the logic probe. Rows denote predictions from the base 72B model, and columns denote predictions from the LoRA-adapted model.

Base 72B	LoRA: preserved	LoRA: violated	Total
preserved	87	4	91
violated	6	3	9
Total	93	7	100

As shown in Table 6, the LoRA-adapted model agrees with the base model on 90 out of 100 cases, yielding an agreement rate of 90.0%. Among them, 87 cases are jointly predicted as `preserved`, and 3 cases are jointly predicted as `violated`. This suggests that after geometry-risk SFT, the model largely preserves the visual logic judgement tendency of the base Qwen2.5-VL-72B model, without exhibiting large-scale prediction drift.

We further summarize two types of disagreement in Table 7.

Table 7: Summary metrics of the logic preservation probe.

Metric	Value
Agreement rate	90.0%
Disagreement rate	10.0%
Boundary relaxation rate	6.0%
Conservative shift rate	4.0%

Here, *boundary relaxation* refers to cases where the base 72B model predicts `violated`, while the LoRA-adapted 72B model predicts `preserved`. This occurs in 6 out of 100 cases, corresponding to a rate of 6.0%. This type of shift is more concerning, as it indicates that the fine-tuned model may become more permissive toward merges that cross functional boundaries in a small number of cases.

In contrast, *conservative shift* refers to cases where the base 72B model predicts `preserved`, while the LoRA-adapted model predicts `violated`. This occurs in 4 out of 100 cases, corresponding to a rate of 4.0%. Compared with boundary relaxation, this type of shift is more conservative: it may reject some acceptable merges, but is less likely to directly introduce erroneous merges.

Overall, this additional experiment indicates that geometry-risk SFT does not substantially overwrite or degrade the original visual logic judgement ability of Qwen2.5-VL-72B. The LoRA-adapted model maintains a 90.0% agreement rate with the base model on the logic probe, suggesting that LoRA mainly learns SC/OCC-related geometry-risk assessment rather than fundamentally changing the model's understanding of semantic and functional part boundaries. However, the 6.0% boundary relaxation rate also suggests that the fine-tuned model may relax part-boundary constraints in a small number of cases. Therefore, in the downstream multi-agent decision pipeline, we do not use the SC/OCC-derived generation risk as the sole decision criterion. Instead, logic correctness is treated as a hard constraint, while geometry-risk signals are used as auxiliary cues for pruning, ranking, and risk assessment.

Inference configuration. At inference time, we load the final GDPO adapter with the Qwen2.5-VL-72B-Instruct backbone using 4-bit quantization and `bfloat16` computation. We use FlashAttention-2, a maximum visual resolution of 262,144 pixels, and a maximum output length of 512 tokens. The same checkpoint and role prompts are used for OmniPart, Hunyuan3D-Part, and MIDI-3D without generator-specific fine-tuning. The numerical ground-truth SC and OCC scores are hidden from the policy during inference. All downstream experiments use the same search budget.

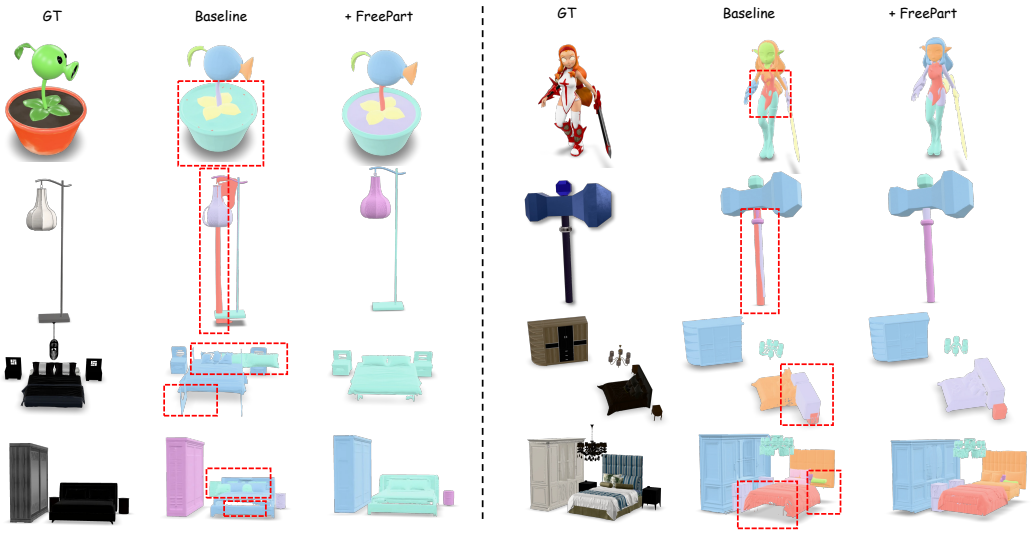


Figure 5: Additional qualitative comparisons across diverse part-level 3D generation pipelines. FreePart reduces fragmented and missing geometry, producing more coherent parts and complete 3D outputs than the corresponding baselines.

Backend-specific configurations. For OmniPart and MIDI-3D, candidate actions are constructed between masks whose boundaries contact or fall within a two-pixel neighborhood. OmniPart uses a mask-area threshold of 2,000 pixels, classifier-free guidance scale of 7.0, and random seed 42.

For Hunyuan3D-Part, the initial mesh partition is obtained using P3-SAM with 20,000 sampled points, 80 prompts, a prompt batch size of 8, and a confidence threshold of 0.95. Candidate actions are derived from shared boundaries between mesh face labels. Hunyuan3D-Part uses 30 inference steps, an octree resolution of 256, four parts per generation batch, and generation seed 2026.

For MIDI-3D, we use a detection threshold of 0.4, a minimum mask-area ratio of 0.001, 50 inference steps, a guidance scale of 7.0, and random seed 42. Polygon-based mask refinement is enabled. Except for the optimized decomposition, all generator weights and backend-specific generation settings are kept identical between the baseline and FreePart.

E Additional Qualitative Comparisons

Figure 5 presents additional qualitative comparisons across image-to-3D object generation, mesh-based part generation, and image-to-3D scene generation. For each comparison, the baseline and FreePart use the same inputs and underlying generators, differing only in the part decomposition. Across diverse objects and scenes, FreePart consistently reduces fragmented geometry, missing parts, and incoherent assemblies, producing more complete parts and structurally coherent 3D outputs. These results further demonstrate the effectiveness and generalizability of decomposition optimization across different part-level 3D generation pipelines.

F User Study

We conducted a blind pairwise user study to evaluate the perceptual quality of the decompositions produced by FreePart. The study included 30 test cases spanning three settings: image-to-3D generation, mesh-to-3D generation, and image-to-3D scene generation. A total of 18 participants completed the study.

For each trial, participants were shown two interactive 3D results generated from the same input using the same downstream generator. The two results differed only in the part decomposition used for generation. To avoid introducing method-name bias, neither the method identities nor the ground-truth result were displayed. Participants could freely rotate and zoom both results before selecting

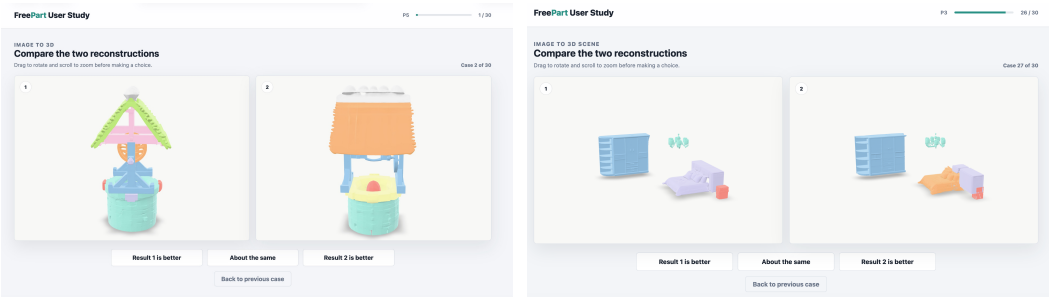


Figure 6: Interface of our blind pairwise user study. Each trial presents two unlabeled interactive 3D results in randomized left–right order. Participants can rotate and zoom both results and select whether Result 1 is better, Result 2 is better, or the two results are about the same. Only one test case is displayed at a time, and participants may return to the previous case to revise their choice.

one of three options: *Result 1 is better*, *Result 2 is better*, or *About the same*. They were instructed to consider both geometric fidelity and overall completeness when making the comparison.

The order of the test cases and the left–right placement of the two methods were independently pseudo-randomized for each participant based on an anonymous participant ID. After each selection, the interface automatically advanced to the next case. Participants could also return to the previous case and revise their response. Upon completion, all responses were stored on the server together with the corresponding participant ID, case identifier, task setting, and randomized method assignment.

Figure 6 shows the evaluation interface. The interface presents one comparison at a time to reduce visual clutter and allows participants to inspect both 3D results under matched viewing conditions.