ARTIREAL: UPGRADING YOUR COARSE ARTICULATED MODELS INTO HIGH-FIDELITY ARTICULATED ASSETS

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality articulated 3D assets are increasingly valuable for robotic interaction, physics-based simulation and digital-twin modeling. Recent articulated-object generators can produce structured models with executable articulation, but still face limitations in fine-grained geometry, realistic appearance, and overall visual coherence. Meanwhile, recent static 3D generation models can synthesize highly realistic assets, yet lack explicit motion properties and have not been fully exploited by existing articulated-object generation pipelines. We introduce **ArtiReal**, a generator-agnostic agentic framework that bridges this gap by upgrading visually coarse articulated assets into high-fidelity ones while preserving their original structure and articulation. To achieve this goal, we first develop an *articulation-aware visual evidence planning* module that actively selects informative observation configurations and constructs coherent and photorealistic object- and part-level references. We then introduce a *structure-preserving adaptive 3D realization* module that adaptively composes geometry and appearance operations to produce high-fidelity parts while preserving articulation constraints. Moreover, an *action-conditioned provenance critic* monitors the entire process, traces failures to their upstream causes, and enables targeted rollback and repair. We evaluate ArtiReal on articulated assets produced by multiple existing generators, with comprehensive comparisons covering visual realism, structural and articulation consistency, efficiency, and systematic ablations. The results demonstrate that ArtiReal consistently improves the visual quality of diverse articulated assets while maintaining their structural and functional validity. *Please refer to the appendix and our anonymous supplementary page for more details and demonstrations.*

1 INTRODUCTION

Articulated objects constitute an important part of everyday human environments, and in recent years, high-quality articulated assets have become an increasingly important foundation for fields such as robotic manipulation, interactive simulation, digital twins, and 3D content creation (Xiang et al., 2020; Jiang et al., 2022; Geng et al., 2022). However, manually creating such assets is costly (Xiang et al., 2020; Wang et al., 2019; Geng et al., 2022), as it requires part-level modeling, kinematic annotation, and joint configuration, motivating a growing body of research on automatic articulated asset generation. Recent methods (Liu et al., 2024b; Su et al., 2025; Gao et al., 2026; Le et al., 2024) generate part geometry and kinematic structures from diverse conditions, including category labels, text, images, point clouds, and existing meshes, through structured generative models or foundation-model-based pipelines. These advances have substantially lowered the barrier to creating structured and interactive 3D content.

Despite this progress, most articulated asset generation methods still focus primarily on part decomposition, kinematic modeling, and structure-conditioned generation, yet remain limited in fine-grained geometry, realistic textures, and overall visual coherence, leaving a substantial gap between their outputs and high-quality real-world assets. In contrast, general-purpose static 3D foundation models (Xiang et al., 2025; Team, 2025; Li et al., 2024) can now generate 3D assets with complex geometry, rich textures, and high-quality materials, yet struggle to explicitly model part identities, joint interfaces, and kinematic constraints. This reveals a clear capability mismatch in the field:

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

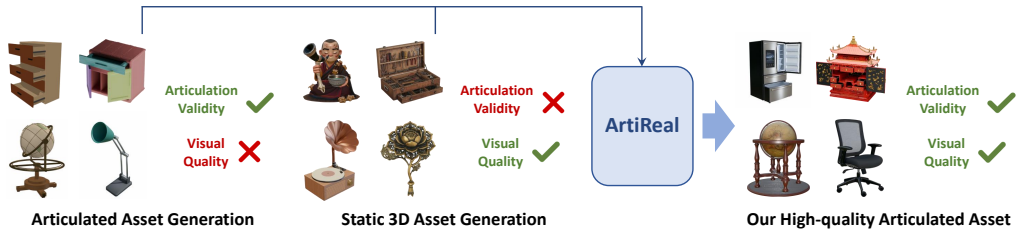

Figure 1: Motivation of ArtiReal. Existing articulated asset generators preserve articulation validity but often lack high visual fidelity, while static 3D generators achieve realistic appearance without explicit articulation. ArtiReal bridges this gap by producing high-quality articulated assets.

existing articulated asset generators excel at producing structurally valid assets with executable articulation, whereas static 3D generators excel at producing visually realistic assets but do not directly support articulated motion. This capability mismatch is illustrated in Fig. 1.

We therefore ask: can the rapidly evolving capabilities of static 3D generation models for high-fidelity asset creation be transferred to visually coarse assets produced by diverse articulated asset generators, while preserving their original part structures and articulation?

To address this problem, we introduce ArtiReal, a generator-agnostic agentic framework for articulation-preserving visual refinement. Given a visually coarse yet structurally valid articulated asset, ArtiReal first performs articulation-aware visual evidence planning: it actively selects informative observation configurations, synthesizes a coherent photorealistic global reference image aligned with the input asset, and derives globally consistent part-level references. It then employs structure-preserving adaptive 3D realization to refine each part's 3D geometry and appearance by adaptively selecting and composing suitable operations according to its structural and visual characteristics, while preserving the original part structure and articulation. Throughout this process, an action-conditioned provenance critic evaluates intermediate results according to the objectives of their corresponding actions, traces detected defects to their responsible upstream stages, and enables targeted repair through rollback. Together, these components form a closed-loop refinement process that coordinates high-fidelity visual generation with explicit articulation constraints.

We conduct extensive experiments on assets produced by diverse articulated asset generators to evaluate ArtiReal from multiple perspectives. Specifically, we compare against several representative refinement pipelines in terms of visual realism, structural and articulation consistency, and human evaluation, and further analyze efficiency and the contribution of key components through dedicated ablation studies. The results demonstrate that ArtiReal consistently improves the visual quality of articulated assets while preserving their structural and functional validity.

Our contributions are summarized as follows: 1) We introduce **ArtiReal**, the first generator-agnostic agentic framework for articulation-preserving visual refinement, bridging the underexplored gap between structurally valid articulated asset generation and high-fidelity static 3D generation. 2) ArtiReal is built around three key technical components for reliable visual refinement: i) **Articulation-aware visual evidence planning**, which turns reference acquisition from passive rendering into active planning over both viewpoint and articulation state, and organizes reference synthesis in a global-to-part manner to provide informative yet appearance-consistent visual evidence for individual parts. ii) **Structure-preserving adaptive 3D realization**, which departs from a one-size-fits-all 3D generator by adaptively orchestrating different 3D operations according to their structural and visual characteristics. iii) **Action-conditioned provenance critic**, which goes beyond local rejection and retry by evaluating results with action-specific objectives, tracing downstream failures to their upstream causes, and enabling targeted rollback and correction. 3) We conduct comprehensive evaluations on assets produced by diverse articulated-object generators, demonstrating that ArtiReal improves visual quality while retaining the functional validity of the source assets.

2 Related Work

Articulated 3D Asset Generation. Early works on articulated objects focused on recovering part structures and motion properties from existing observations, including motion-aware part decomposition (Wang et al., 2019), disentangled shape–articulation representations (Mu et al., 2021), and

articulated digital-twin reconstruction (Jiang et al., 2022). More recent methods directly synthesize articulated objects and have progressively improved controllability, conditioning diversity, and geometric fidelity. Representative examples include structured generative models (Lei et al., 2023; Liu et al., 2024b; Leboutet et al., 2024), image-conditioned methods (Liu et al., 2024a; Wu et al., 2025), and recent approaches that generate more detailed part geometry and meshes (Su et al., 2025; Gao et al., 2025). Procedural and foundation-model-based approaches further extend articulated asset creation through scalable synthesis (Sun et al., 2025) and language- or agent-driven modeling (Gao et al., 2026; Zhou et al., 2026; Mandi et al., 2024; Le et al., 2024; Wang et al., 2026). Despite these advances, most articulated asset generation methods still place greater emphasis on part geometry, structural composition, and kinematic correctness than on high-fidelity visual appearance. Several recent works have begun to jointly consider articulation and appearance (Wei et al., 2022; Guo et al., 2025), but they mainly focus on reconstructing articulated objects from visual observations, and their ability to produce high-fidelity visual appearance remains limited.

High-Fidelity Static 3D Generation. Static 3D generation has advanced rapidly with the development of large-scale image diffusion and 3D generative models. Early approaches leverage powerful 2D diffusion priors to optimize 3D representations, progressively improving geometry and appearance quality (Poole et al., 2022; Lin et al., 2023; Chen et al., 2023b). Subsequent works move toward scalable and feed-forward generation from image or text conditions (Zhao et al., 2023; Hong et al., 2023; Liu et al., 2024c; Tang et al., 2024), while recent native 3D generative models further push the fidelity of explicit geometry, textures, and materials (Xiang et al., 2024; Li et al., 2024; Chen et al., 2025; Xiang et al., 2025; Team, 2025). In parallel, specialized appearance-generation methods have substantially improved texture consistency and material quality (Chen et al., 2023a; Zeng et al., 2024; Cheng et al., 2025; Xiong et al., 2024). Despite their strong visual generation capabilities, these methods typically treat an object as a static entity and do not explicitly maintain part identities, joint relations, or kinematic constraints, making their direct application to articulated assets challenging.

Agentic Generation Systems. With the rapid development of large language and vision-language models, recent works increasingly formulate content generation as an agentic process rather than a single feed-forward prediction. Early systems employ language models to decompose and execute complex 3D modeling tasks through procedural tools (Sun et al., 2023; Yang et al., 2024), while program-based representations further improve the controllability of automated scene generation (Zhang et al., 2025). In parallel, multimodal generation agents introduce iterative verification and self-refinement: Yang et al. (2023) repeatedly evaluates generated images and revises generation prompts, while Zhenyu et al. (2024) orchestrates specialized generation and editing tools through planning and step-wise self-correction. More recent systems extend this paradigm to articulated objects and complex 3D environments using collaborative agents, tool execution, and critic-guided refinement (Le et al., 2024; Gao et al., 2026; Xia et al., 2026). These developments inspire us to adopt an agentic, closed-loop paradigm to address the problem of articulation-preserving visual refinement, enabling adaptive orchestration of heterogeneous 3D tools and iterative correction of intermediate failures.

3 METHOD

3.1 PROBLEM SETUP AND AGENTIC OVERVIEW

Problem Definition. We consider an articulated asset represented by a set of rigid parts $\mathcal{V} = \{v_i\}_{i=1}^{N}$ and a kinematic graph $\mathcal{K} = (\mathcal{V}, \mathcal{E})$. Each part is represented as $v_i = (\mathcal{S}_i, \mathcal{A}_i, \boldsymbol{\rho}_i)$, where $\mathcal{S}_i$ denotes the part geometry, $\mathcal{A}_i$ denotes its appearance, and $\boldsymbol{\rho}_i$ denotes its kinematic attributes, including the reference transformation and articulation parameters associated with the part. Specifically, $\mathcal{S}_i$ is typically represented as a triangle mesh, while $\mathcal{A}_i$ describes the corresponding color, texture, and material attributes. Given an input asset $\mathcal{A}^0 = (\{\mathcal{S}_i^0, \mathcal{A}_i^0, \boldsymbol{\rho}_i^0\}_{i=1}^{N}, \mathcal{K}^0)$, whose articulation structure is assumed to be valid but whose geometry and appearance may be visually incomplete or low quality, our objective is to produce a refined asset $\mathcal{A}^\star = (\{\mathcal{S}_i^\star, \mathcal{A}_i^\star, \boldsymbol{\rho}_i^0\}_{i=1}^{N}, \mathcal{K}^0)$. The refinement is allowed to update only the geometry and appearance of each part while preserving the original kinematic structure and motion attributes. Meanwhile, the refined parts are expected to remain mutually compatible in geometry and coherent in appearance, yielding a visually high-quality articulated asset without compromising its original articulation.

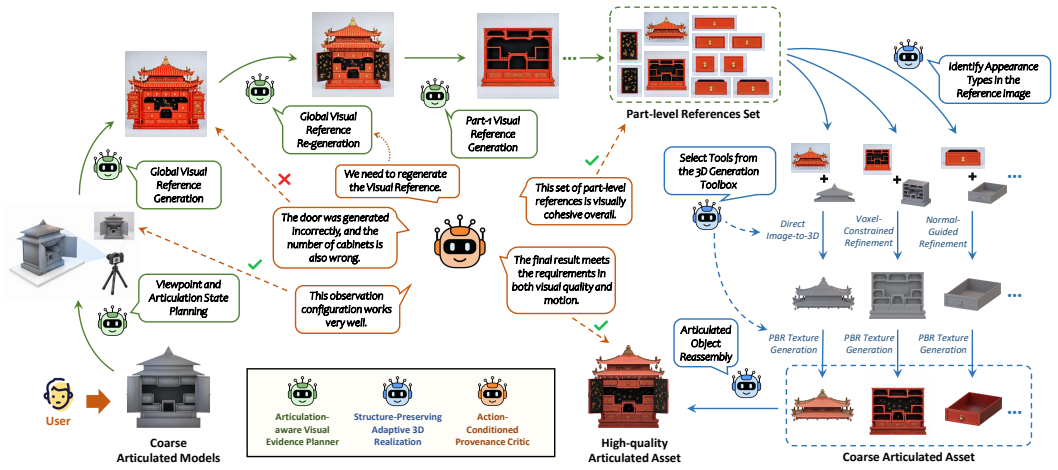

Figure 2: Overview of **ArtiReal**. Starting from a coarse articulated model, ArtiReal plans informative observations, synthesizes coherent object- and part-level visual references, adaptively realizes high-fidelity 3D parts with structure-preserving tool selection, and reassembles them into a high-quality articulated asset under critic-guided refinement.

ArtiReal Overview. To achieve the aforementioned objective, we develop *ArtiReal*, a generator-agnostic agentic framework for refining visually coarse articulated 3D assets into high-fidelity ones. ArtiReal achieves this goal through the coordinated interaction of three core modules: an articulation-aware visual evidence planner, a structure-preserving adaptive 3D realization module, and an action-conditioned provenance critic. An overview of ArtiReal is illustrated in Fig. 2.

Specifically, given a visually coarse articulated asset $\mathcal{A}^0$, the visual evidence planner first selects an informative viewpoint and articulation state, and synthesizes a photorealistic global reference image $\mathcal{I}^{\text{obj}}$ that captures the complete object with enriched geometry and appearance. Based on this global reference, it further generates a set of part-level references $\mathcal{I}^{\text{part}} = \{\mathcal{I}_i^{\text{part}}\}_{i=1}^{N}$, where each reference remains aligned with the corresponding coarse part and consistent with the global visual style. The 3D realization module then takes these part-level references as visual guidance and adaptively selects and composes suitable operations from a predefined *3D Generation Toolbox* to refine each part's geometry and appearance while preserving the original articulation structure. Throughout the process, the critic evaluates intermediate results and, when necessary, traces failures to their responsible upstream stages to trigger targeted rollback and revision. This iterative refinement continues until a visually high-quality and structurally consistent articulated asset is obtained.

3.2 Articulation-Aware Visual Evidence Planner

To improve the geometry and appearance of a visually coarse articulated asset, the first step is to collect reliable and sufficiently informative visual evidence. We therefore design an *Articulation-Aware Visual Evidence Planner* to adaptively determine the observation conditions and construct visually coherent references for subsequent 3D refinement. To enable this process, we equip the agent with three types of executable actions: observation configuration, global reference synthesis, and part-level reference synthesis. The observation configuration action jointly adjusts the rendering viewpoint and the articulation state, while the latter two actions generate photorealistic references at the object and part levels, respectively.

Specifically, given the input articulated asset $\mathcal{A}^0$, the agent selects an observation configuration $\omega = (\mathbf{c}, \boldsymbol{\theta})$, where $\mathbf{c}$ denotes the camera configuration and $\boldsymbol{\theta}$ denotes the articulation state. Rendering $\mathcal{A}^0$ under ω produces an observation $\mathcal{O}$, which serves as the structural condition for synthesizing a global reference $\mathcal{I}^{\text{obj}}$. Then, for each part v_i, the agent uses the global reference together with the corresponding coarse geometry $\mathcal{S}_i^0$ to generate a part-level reference $\mathcal{I}_i^{\text{part}}$. During this process, the agent evaluates whether the current observation provides sufficient evidence for each part. If a part is heavily occluded or its geometry and appearance are not adequately revealed, the agent plans a part-specific observation configuration ω_i and renders the asset again under the revised viewpoint

and articulation state. The resulting observation is then used to synthesize a revised global reference and a more informative part-level reference for v_i.

Observation Configuration. The observation configuration action seeks a camera–articulation configuration that is informative for subsequent reference synthesis. In particular, a desirable configuration should (i) provide a coherent and semantically plausible object-level view, and (ii) expose relevant parts and articulation cues while reducing self-occlusion. On its first invocation, the system analyzes the articulated asset to determine a canonical front direction and an initial articulation state. Around this reference configuration, it samples candidate camera viewpoints and articulation states, renders them using the original asset, and filters them according to deterministic geometric criteria such as part visibility, object completeness, and joint observability. A vision-language model then selects the most informative configuration from the geometry-valid candidates. On subsequent invocations, the system first revisits the remaining valid candidates to determine whether any better reveals the currently under-observed parts. If none is suitable, the model proposes a refinement direction, around which additional candidates are sampled, filtered, and selected.

Global Reference Synthesis. The goal of global reference synthesis is to enrich the coarse articulated asset with realistic geometric details, textures, and materials while preserving its original part structure and articulation configuration, and maintaining appearance coherence across parts. Specifically, the agent first specifies an object-level appearance design, including the materials, colors, geometric details, and the visual treatment of inter-part interfaces. This design serves as the appearance condition, while the coarse asset rendered under the observation configuration provides the structural condition. Conditioned on both, an image generation model synthesizes a photorealistic global reference $\mathcal{I}^{\mathrm{obj}}$. The appearance condition can also be derived from an existing image, enabling appearance transfer or consistent regeneration under a revised observation configuration.

Part-Level Reference Synthesis. Given the global reference $\mathcal{I}^{\mathrm{obj}}$, the agent generates a photorealistic reference for each part v_i, using the global image as an appearance prior to maintain object-level visual consistency. For each part, the image generation model is additionally conditioned on its location within the full object for semantic and spatial context, together with an isolated rendering of the corresponding coarse geometry $\mathcal{S}_i^0$ as the structural condition.

3.3 Structure-Preserving Adaptive 3D Realization

Having obtained part-level references from the visual evidence planner, the next step is to translate this visual evidence into high-fidelity 3D parts while preserving their compatibility with the original articulated asset. A key challenge is that different parts may require substantially different refinement strategies. For example, articulation-critical components should preserve their overall shape and spatial extent, whereas small decorative components can tolerate larger geometric changes in exchange for improved visual fidelity. Similarly, some references mainly require geometric refinement, while others are dominated by complex textures or surface materials. Therefore, rather than relying on a single 3D generator, ArtiReal adaptively selects and composes operations from a *3D Generation Toolbox* according to the structural and visual characteristics of each part.

For each part v_i, the agent jointly analyzes its role in the articulated object and the corresponding part-level reference $\mathcal{I}_i^{\mathrm{part}}$. In particular, it considers factors such as articulation sensitivity, geometric discrepancy from the coarse geometry $\mathcal{S}_i^0$, geometric-detail complexity, and appearance characteristics. Based on these properties, the agent determines how much geometric freedom can be safely introduced and selects an appropriate combination of geometry refinement, appearance realization, and mesh maintenance operations. This adaptive strategy allows articulation-critical parts to remain strongly constrained by the original geometry $\mathcal{S}_i^0$, while visually complex but structurally less sensitive parts can be reconstructed with greater flexibility. Specifically, the 3D Generation Toolbox is organized into three categories of operations, as summarized in Tab. 1.

3.4 Action-Conditioned Provenance Critic

A multi-stage generation pipeline poses a distinct challenge for quality control: the validity of an intermediate result cannot be assessed using a single fixed criterion, and the stage where a failure becomes observable does not necessarily coincide with the stage that caused it. We therefore intro-

Table 1: Overview of the 3D Generation Toolbox. The agent adaptively selects and composes operations according to the structural and visual characteristics of each part.

Category	Operation	When to Use	Realization
Geometry	Voxel-Constrained Refinement	Global shape and scale should be preserved, while moderate geometric changes are required.	Uses the coarse part as a voxel condition to preserve global structure while introducing moderate geometric details.
	Normal-Guided Refinement	Global geometry is reliable, but fine and high-frequency surface structures need to be recovered.	Refines local surface geometry under normal guidance to reproduce fine-scale surface details.
	Direct Image-to-3D	The part is articulation-insensitive and contains highly complex geometry.	Reconstructs the part directly from its visual reference with greater geometric freedom.
Appearance	PBR Texture Generation	General materials or spatially varying surface appearance need to be synthesized.	Generates coherent PBR materials and textures conditioned on the part-level visual reference.
	Texture Projection	The surface is relatively flat and contains intricate or highly specific visual patterns.	Projects image-space appearance onto the corresponding surface for faithful texture reproduction.
Maintenance	Watertight Repair	The generated mesh contains holes or non-watertight regions.	Repairs mesh defects to recover a watertight surface.
	Mesh Simplification	Polygon density is unnecessarily high relative to geometric complexity.	Removes redundant faces while preserving the represented geometry.

duce an **Action-Conditioned Provenance Critic**, which evaluates intermediate results throughout the generation process according to the intended objective of the action that produced them.

Concretely, after each action, the critic inspects its output together with the corresponding action intent and determines whether the intended requirement has been satisfied. If the result is accepted, the agent proceeds to the next step. Otherwise, instead of simply rejecting the current output, the critic analyzes the observed defect in the context of the preceding generation trajectory. This is important because an error exposed at a later stage may originate from an earlier decision. For instance, an unsatisfactory 3D reconstruction may be caused by an inappropriate 3D generation operation, but it may also stem from an uninformative viewpoint or an inaccurate visual reference generated upstream. The critic therefore performs *provenance-aware diagnosis* to identify the action most likely responsible for the failure and provides a corresponding corrective suggestion. The agent then rolls back to the attributed stage and revises the responsible decision, rather than blindly modifying the latest output.

Retrospective Adjudication. Moreover, since downstream evidence may occasionally contradict earlier critic decisions, we introduce a lightweight *retrospective adjudication* step to determine whether an earlier result was incorrectly accepted or a later diagnosis is overly aggressive before performing rollback. This prevents inconsistent assessments across stages from causing repeated correction loops and unnecessary computational cost.

4 Experiment

We conduct comprehensive experiments to evaluate the effectiveness of ArtiReal across diverse articulated assets. We first describe the experimental setup in Sec. 4.1, followed by the main comparisons and analysis in Sec. 4.2. We then evaluate refinement efficiency in Sec. 4.3. Finally, Sec. 4.4 analyzes the contribution of the key designs through ablation studies.

4.1 Experimental Setup

Test Dataset. To evaluate the generalizability of ArtiReal across different articulated asset generators, we construct a test set using assets produced by four representative methods: ArtFormer, Arti-PG, LAM, and ArtiCraft. For each source method, we collect 100 articulated assets and, according to the object categories supported by each method, select the generated samples to cover as diverse a range of categories as possible.

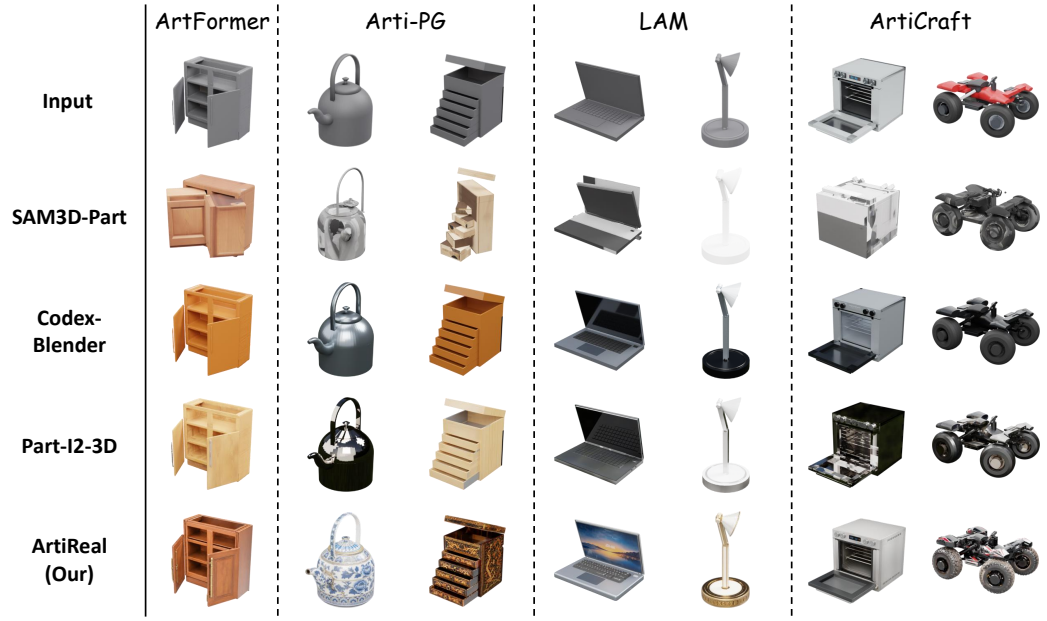

Figure 3: Qualitative comparison of different visual refinement methods across articulated assets generated by diverse source models.

Baselines. Since no prior method directly targets visual refinement of existing articulated assets while preserving their articulation structure, we construct three representative baseline pipelines using recent image generation, image-to-3D reconstruction, and general-purpose agentic editing tools. All baselines share the same refinement objective: improving geometry and appearance while preserving the original part identities, composition, and articulation configuration. **Mask-Guided Part Reconstruction (SAM3D-Part)** first generates a photorealistic full-object image, extracts the image region of each articulated part, and reconstructs each part independently with SAM 3D Objects (Team et al., 2025), followed by alignment and reassembly under the original articulation structure. **Agentic Blender Editing (Codex-Blender)** uses a strong general-purpose agent (OpenAI, 2025) equipped with Blender to directly refine part geometry, materials, textures, and colors while preserving the original articulation structure. **Independent Part Regeneration (Part-I2-3D)** generates a photorealistic full-object reference, segments it into part-level images, and independently reconstructs each part using TRELLIS.2 (Xiang et al., 2025) and Hunyuan3D 2.1 (Team, 2025), followed by alignment and reassembly according to the source articulation structure.

Evaluation Metrics. To evaluate the visual quality of the refined assets, we use FID and KID to measure the distribution gap between rendered results and reference-object renderings, where lower values indicate better realism. We further report CLIP-I and BLIP scores to assess how well the refined assets preserve the semantic and visual characteristics of the source objects. To measure structural and articulation consistency, we use Motion-CD to evaluate geometric deviation under articulation, AID to assess the consistency of articulated structure, and AOR to measure the extent of part overlap and interaction inconsistency, where lower values indicate better preservation. Finally, we conduct a human study from two complementary perspectives: visual quality and realism, and the overall consistency of part structure and articulation, with each aspect rated on a five-point scale.

Implementation Details. We use GPT-5.6 Sol (OpenAI, 2026a) as the agent backbone for planning, tool selection, and critic reasoning, and GPT-Image-2 (OpenAI, 2026b) for global and part-level reference synthesis. For the 3D Generation Toolbox, voxel-constrained geometry refinement is implemented by adapting TRELLIS.2 (Xiang et al., 2025) to condition its geometry realization on the coarse voxelized part. For normal-guided surface refinement, the agent first infers multi-view normal maps from the part-level reference to capture fine geometric details, and then optimizes the mesh through continuous remeshing (Palfinger, 2022) under the predicted normal guidance. Direct image-to-3D generation is performed using Hunyuan3D 2.1 (Team, 2025), while Hunyuan3D-Paint 2.1 (Team, 2025) is used for PBR texture generation. During observation configuration, we sample

Table 2: Quantitative comparison of different visual refinement methods across articulated assets generated by different source methods.

Source	Method	Visual Fidelity				Structural & Articulation Consistency			Human Evaluation	
		FID↓	KID×100 ↓	CLIP-I↑	BLIP↑	Motion-CD↓	AID↓	AOR↓	Realism↑	Consistency↑
ArtFormer	Input	130.191	13.618	0.905	0.403	–	–	–	2.37	–
	SAM3D-Part	115.048	10.357	0.929	0.460	0.073	0.296	0.012	3.72	3.53
	Codex-Blender	125.164	12.121	0.934	0.435	**0.000**	0.069	**0.000**	3.58	4.41
	Part-I2-3D	124.852	12.535	0.935	0.444	**0.000**	**0.023**	**0.000**	3.31	**4.67**
	ArtiReal	**108.664**	**9.339**	**0.960**	**0.504**	**0.000**	0.043	**0.000**	**4.16**	4.59
Arti-PG	Input	73.722	1.903	0.873	0.301	–	–	–	2.81	–
	SAM3D-Part	69.365	1.092	0.909	0.277	0.034	0.294	0.034	3.44	3.50
	Codex-Blender	63.136	0.994	0.908	0.323	**0.000**	0.086	0.021	3.92	4.71
	Part-I2-3D	62.477	0.876	0.908	0.288	0.001	0.077	**0.015**	3.89	4.46
	ArtiReal	**62.056**	**0.780**	**0.923**	**0.520**	**0.000**	**0.046**	**0.015**	**4.67**	**4.75**
LAM	Input	74.241	2.229	0.859	0.269	–	–	–	2.44	–
	SAM3D-Part	68.760	1.524	0.891	0.249	0.508	0.304	0.013	3.39	3.15
	Codex-Blender	62.276	1.437	0.890	0.286	**0.001**	**0.008**	0.016	3.70	4.71
	Part-I2-3D	64.257	1.221	0.890	0.275	0.078	0.012	0.017	3.61	4.37
	ArtiReal	**58.952**	**1.025**	**0.909**	**0.585**	**0.001**	0.032	**0.003**	**4.38**	**4.76**
ArtiCraft	Input	77.580	2.215	0.859	0.261	–	–	–	2.73	–
	SAM3D-Part	67.052	0.930	0.896	0.299	0.052	0.253	0.013	3.64	3.07
	Codex-Blender	**63.641**	1.979	0.893	0.314	0.002	0.018	0.004	4.09	4.33
	Part-I2-3D	66.596	0.940	0.896	0.311	**0.000**	**0.003**	**0.002**	3.83	**4.80**
	ArtiReal	63.987	**0.776**	**0.911**	**0.599**	**0.000**	0.066	**0.002**	**4.57**	4.69

five candidate camera–articulation configurations in each round. We use PartNet-Mobility (Xiang et al., 2020) as the reference set of articulated assets and render them using the same protocol.

4.2 Results & Analysis

We compare ArtiReal with the three baseline pipelines both quantitatively and qualitatively, as shown in Tab. 2 and Fig. 3. Specifically:

Visual Fidelity. ArtiReal achieves the best overall visual quality across nearly all visual fidelity metrics, which is also clearly reflected in the qualitative comparisons in Fig. 3. In contrast, the baseline methods exhibit several characteristic failure modes. SAM3D-Part often suffers from severe realism degradation due to erroneous geometry generation of individual parts, while Codex-Blender and Part-I2-3D generally lack rich geometric details and tend to produce overly simple or even incorrect textures, resulting in insufficient visual realism. By constructing globally coherent visual references and adaptively selecting suitable 3D realization strategies for different parts, ArtiReal produces richer geometric details, more realistic textures and materials, and stronger object-level visual coherence.

Structural and Articulation Consistency. In terms of structural and articulation consistency, ArtiReal achieves comparable performance to Codex-Blender and Part-I2-3D, with all three methods largely preserving the original part structure and articulation behavior. By contrast, SAM3D-Part performs noticeably worse in both the quantitative evaluation and qualitative results, frequently introducing part deformation, scale inconsistency, and misalignment after reconstruction. This is mainly because SAM3D-Part reconstructs each segmented part independently without explicitly constraining it to preserve the original geometry and kinematic compatibility.

4.3 Efficiency Analysis

We further evaluate the efficiency of ArtiReal in reducing the manual effort required for articulated asset refinement. We record the end-to-end runtime of ArtiReal over the test set and additionally select 50 cases for manual refinement. For each selected case, participants are provided with the same source articulated asset and asked to improve its geometry and appearance using available 3D modeling, image generation, and editing tools while preserving the original part structure and articulation. On average, ArtiReal completes the refinement of one asset in **27 minutes**, compared with **147 minutes** required for manual refinement. This corresponds to a $5.45\times$ speedup in completion time, demonstrating that ArtiReal substantially reduces the human effort required to produce visually refined articulated assets.



Figure 4: Qualitative comparison of different ablation variants of ArtiReal.

Table 3: Ablation study of the key components in ArtiReal.

Variant	Visual Fidelity				Structural & Articulation Consistency			Human Evaluation	
	FID↓	KID×100 ↓	CLIP-I↑	BLIP↑	Motion-CD↓	AID↓	AOR↓	Realism↑	Consistency↑
w/o Observation Planning	73.433	2.239	0.901	0.487	**0.000**	0.051	**0.004**	4.26	4.69
w/o Global Appearance Reference	75.916	2.457	0.890	0.437	0.001	0.054	**0.004**	4.08	4.57
w/o Critic	80.218	2.781	0.875	0.420	0.004	0.087	0.005	3.77	4.59
Local Critic	72.109	2.243	0.889	0.504	**0.000**	**0.036**	0.005	4.37	4.61
Full ArtiReal	**71.388**	**2.176**	**0.928**	**0.531**	**0.000**	0.041	**0.004**	**4.45**	**4.72**

4.4 Ablation Studies

We conduct ablation studies to examine the contribution of the key designs in ArtiReal. First, to evaluate the importance of active observation selection, we remove the observation planning mechanism (*w/o Observation Planning*) and directly use the initial camera viewpoint and articulation state for subsequent reference synthesis. Second, to study the role of hierarchical visual evidence construction, we remove the global appearance reference (*w/o Global Appearance Reference*) and independently synthesize the photorealistic reference for each part from its coarse rendering, without conditioning on a coherent full-object reference. We further investigate the critic design at two levels. In *w/o Critic*, all intermediate evaluations and corrective feedback are disabled, and the pipeline proceeds without iterative revision. In *Local Critic*, the critic retains failure detection and retry, but correction is restricted to the current action without tracing the defect to an upstream source.

As shown in Tab. 3, the full ArtiReal consistently achieves the best visual-fidelity results, while the structural and articulation consistency metrics remain comparable across different variants. The qualitative results in Fig. 4 further reveal distinct failure modes for different ablations. Without observation planning, some internal components are poorly exposed during reference synthesis, leading to abnormal or inconsistent textures. Removing the global appearance reference often causes cross-part style inconsistency or semantically incorrect part textures. Without the critic, the instability of single-pass generation can directly propagate to the final result, producing substantial appearance noise and visual artifacts. Although the local critic can partially alleviate such failures through repeated retry, its visual quality remains inferior to the full system.

We additionally analyze the effect of *retrospective adjudication*. Across the evaluation set, 73 cases involved correction cycles exceeding three rounds, while retrospective adjudication helps resolve inconsistent critic decisions and reduce unnecessary regeneration.

5 Conclusion

In this work, we introduced **ArtiReal**, a generator-agnostic agentic framework that upgrades visually coarse articulated models into high-fidelity assets while preserving their original structure and articulation. ArtiReal combines articulation-aware visual evidence planning, structure-preserving adaptive 3D realization, and action-conditioned provenance critic in a closed-loop refinement process. Experiments on assets from diverse articulated object generators show that ArtiReal substantially improves visual quality while maintaining structural and functional validity. We believe this work provides a general paradigm for integrating the rapidly advancing capabilities of static 3D generation with the unique structural requirements of articulated assets.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing of the manuscript. Generative AI models are also used as explicit components of the proposed method and experimental baselines, as described in the main paper; these uses are part of the evaluated systems rather than undisclosed authoring assistance. We did not rely on generative AI to independently determine experimental results or scientific conclusions. All AI-assisted content was manually reviewed, revised, and verified by the authors. In particular, factual claims and literature references were checked against their original sources, and experimental results were obtained and validated through the evaluation procedures described in the paper. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023a.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22246–22256, October 2023b.

Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, Liang Pan, Dahua Lin, and Ziwei Liu. 3dtopia-xl: High-quality 3d pbr asset generation via primitive diffusion. In *CVPR*, 2025.

Wei Cheng, Juncheng Mu, Xianfang Zeng, Xin Chen, Anqi Pang, Chi Zhang, Zhibin Wang, Bin Fu, Gang Yu, Ziwei Liu, et al. Mvpaint: Synchronized multi-view diffusion for painting anything 3d. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 585–594, 2025.

Daoyi Gao, Yawar Siddiqui, Lei Li, and Angela Dai. Meshart: Generating articulated meshes with structure-guided transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Yipeng Gao, Yunhao Ge, Peilin Cai, Daniel Seita, and Laurent Itti. Lam: Language articulated object modelers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. *arXiv preprint arXiv:2211.05272*, 2022.

Junfu Guo, Yu Xin, Gaoyi Liu, Kai Xu, Ligang Liu, and Ruizhen Hu. Articulatedgs: Self-supervised digital twin modeling of articulated objects using 3d gaussian splatting. *arXiv preprint arXiv:2503.08135*, 2025.

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. *arXiv preprint arXiv:2311.04400*, 2023.

Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *arXiv preprint arXiv:2202.08227*, 2022.

Long Le, Jason Xie, William Liang, Hung-Ju Wang, Yue Yang, Yecheng Jason Ma, Kyle Vedder, Arjun Krishna, Dinesh Jayaraman, and Eric Eaton. Articulate-anything: Automatic modeling of articulated objects via a vision-language foundation model. *arXiv preprint arXiv:2410.13882*, 2024.

Quentin Leboutet, Nina Wiedemann, Zhipeng Cai, Michael Paulitsch, and Kai Yuan. Midgard: Modular interpretable diffusion over graphs for articulated designs. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 1556–1585. Curran Associates, Inc., 2024. doi: 10.52202/

079017-0050. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0318de478e18308a5f64297f618299d3-Paper-Conference.pdf.

Jiahui Lei, Congyue Deng, Bokui Shen, Leonidas Guibas, and Kostas Daniilidis. Nap: Neural 3d articulation prior, 2023. URL https://arxiv.org/abs/2305.16315.

Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner, 2024.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation, 2023. URL https://arxiv.org/abs/2211.10440.

Jiayi Liu, Denys Iliash, Angel X Chang, Manolis Savva, and Ali Mahdavi-Amiri. SINGAPO: Single image controlled generation of articulated parts in object. *arXiv preprint arXiv:2410.16499*, 2024a.

Jiayi Liu, Hou In Ivan Tam, Ali Mahdavi-Amiri, and Manolis Savva. Cage: Controllable articulation generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17880–17889, 2024b.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024c.

Zhao Mandi, Yijia Weng, Dominik Bauer, and Shuran Song. Real2code: Reconstruct articulated objects via code generation, 2024. URL https://arxiv.org/abs/2406.08474.

Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan L. Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-SDF: learning disentangled signed distance functions for articulated shape representation. pp. 12981–12991, 2021.

OpenAI. Introducing codex. https://openai.com/index/introducing-codex/, May 2025. Accessed: 2026-09-20.

OpenAI. GPT-5.6 Sol, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. OpenAI API model documentation, accessed September 20, 2026.

OpenAI. GPT-Image-2, 2026b. URL https://developers.openai.com/api/docs/models/gpt-image-2. OpenAI API model documentation, accessed September 20, 2026.

Werner Palfinger. Continuous remeshing for inverse rendering. *Computer Animation and Virtual Worlds*, 33, 07 2022. doi: 10.1002/cav.2101.

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

Jiayi Su, Youhe Feng, Zheng Li, Jinhua Song, Yangfan He, Botao Ren, and Botian Xu. Artformer: Controllable generation of diverse 3d articulated objects, 2025. URL https://arxiv.org/abs/2412.07237.

Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 3d-gpt: Procedural 3d modeling with large language models. *arXiv preprint arXiv:2310.12945*, 2023.

Jianhua Sun, Yuxuan Li, Jiude Wei, Longfei Xu, Nange Wang, Yining Zhang, and Cewu Lu. Artipg: A toolbox for procedurally synthesizing large-scale and diverse articulated objects with rich annotations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6396–6405, 2025.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei 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. 2025. URL https://arxiv.org/abs/2511.16624.

Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025.

Penghao Wang, Siyuan Xie, Hongyu Yan, Xianghui Yang, Jingwei Huang, Chunchao Guo, and Jiayuan Gu. Artllm: Generating articulated assets via 3d llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34281–34291, 2026.

Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. *IEEE Conference on Computer Vision and Pattern*, pp. to appear, 2019.

Fangyin Wei, Rohan Chabra, Lingni Ma, Christoph Lassner, Michael Zollhoefer, Szymon Rusinkiewicz, Chris Sweeney, Richard Newcombe, and Mira Slavcheva. Self-supervised neural articulated shape and appearance models. In *Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ruqi Wu, Xinjie Wang, Liu Liu, Chunle Guo, Jiaxiong Qiu, Chongyi Li, Lichao Huang, Zhizhong Su, and Ming-Ming Cheng. Dipo: Dual-state images controlled articulated object generation powered by diverse data. In *Advances in Neural Information Processing Systems 39 (NeurIPS 2025)*, 2025.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, Shuran Song, and Fangyin Wei. Sage: Scalable agentic 3d scene generation for embodied ai. *arXiv preprint arXiv:2602.10116*, 2026.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

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. *Tech report*, 2025.

Bojun Xiong, Jialun Liu, Jiakui Hu, Chenming Wu, Jinbo Wu, Xing Liu, Chen Zhao, Errui Ding, and Zhouhui Lian. Texgaussian: Generating high-quality pbr material via octree-based 3d gaussian splatting. *arXiv preprint arXiv:2411.19654*, 2024.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16227–16237, June 2024.

Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2img: Iterative self-refinement with gpt-4v(ision) for automatic image design and generation. *arXiv preprint arXiv:2310.08541*, 2023.

Xianfang Zeng, Xin Chen, Zhongqi Qi, Wen Liu, Zibo Zhao, Zhibin Wang, Bin Fu, Yong Liu, and Gang Yu. Paint3d: Paint anything 3d with lighting-less texture diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4252–4262, 2024.

Yunzhi Zhang, Zizhang Li, Matt Zhou, Shangzhe Wu, and Jiajun Wu. The scene language: Representing scenes with programs, words, and embeddings. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24625–24634, 2025.

Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, BIN FU, Tao Chen, Gang YU, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xmxgMij3LY.

Wang Zhenyu, Li Aoxue, Li Zhenguo, and Liu Xihui. Genartist: Multimodal llm as an agent for unified image generation and editing. *arXiv preprint arXiv:2407.05600*, 2024.

Matt Zhou, Ruining Li, Xiaoyang Lyu, Zhaomou Song, Zhening Huang, Chuanxia Zheng, Christian Rupprecht, Andrea Vedaldi, and Shangzhe Wu. Articraft: An agentic system for scalable articulated 3d asset generation. *arXiv preprint arXiv:2605.15187*, 2026.

A Implementation Details of the 3D Generation Toolbox

The 3D Generation Toolbox contains a set of geometry, appearance, and mesh-maintenance operations with different degrees of geometric freedom. For each part, the agent selects and composes these operations according to its articulation sensitivity, geometric complexity, appearance characteristics, and the discrepancy between the coarse geometry and the corresponding visual reference. Before applying any geometry-generation operation, we normalize the coarse part into a canonical coordinate frame while recording its original scale, orientation, and transformation. After refinement, the resulting geometry is transformed back to the original part coordinate system so that the kinematic structure of the articulated asset remains unchanged.

A.1 Voxel-Constrained Refinement

Voxel-Constrained Refinement is used when the global shape, scale, and spatial occupancy of a part should remain close to the original asset while moderate geometric enhancement is allowed. We adapt the geometry realization stage of TRELLIS.2 (Xiang et al., 2025) by introducing a voxelized representation of the coarse geometry $\mathcal{S}_i^0$ as an explicit structural condition.

Specifically, the coarse part is first normalized to a unit bounding box while preserving its aspect ratio, and voxelized into a 128^3 occupancy grid. The voxel grid is then used to constrain the spatial support of the generated geometry, encouraging the refined part to remain close to the original shape and scale while allowing local geometric variations such as smoother boundaries, rounded corners, shallow holes, and low-to-mid-frequency surface details.

To incorporate this structural constraint, we modify the original TRELLIS.2 inference pipeline by bypassing its sparse-structure generation stage and directly reusing the subsequent shape structured-latent realization stage. Concretely, the occupied coordinates extracted from the voxelized coarse part are used in place of the sparse structure predicted by TRELLIS.2, while the part-level visual reference provides the image condition for shape realization. The resulting structured latent is then decoded using the original TRELLIS.2 geometry decoder.

A.2 Normal-Guided Surface Refinement

Normal-Guided Refinement is used when the coarse global geometry is reliable but fine-scale surface structures are missing. Instead of regenerating the part geometry, this operation directly deforms the original mesh under predicted surface-normal guidance.

We first render the coarse part from six canonical viewpoints corresponding to the positive and negative directions of the three principal axes. Conditioned on the part-level visual reference $\mathcal{I}_i^{\text{part}}$ and the corresponding coarse renderings, the agent predicts a normal map for each visible view. These multi-view normal maps provide complementary geometric cues for local surface structures such as grooves, engravings, shallow reliefs, seams, and other high-frequency details.

The input mesh is then optimized through continuous remeshing (Palfinger, 2022). At each iteration, the current mesh is differentiably rendered into the selected views, and the rendered normals are compared with the predicted normal maps. We optimize the vertex positions using a combination of a normal-consistency loss, a Laplacian smoothness regularizer, an edge-length regularizer, and a displacement penalty that discourages large deviations from the original coarse surface:

$$\mathcal{L} = \lambda_{\text{n}}\mathcal{L}_{\text{normal}} + \lambda_{\text{lap}}\mathcal{L}_{\text{lap}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}} + \lambda_{\text{disp}}\mathcal{L}_{\text{disp}}. \tag{1}$$

The mesh is optimized for 500 iterations using Adam with a learning rate of 1×10^{-3}. Continuous remeshing is performed every 20 iterations to maintain approximately uniform triangle quality and to introduce additional vertices where fine geometric details are required. Since deformation is directly regularized against the original mesh, this operation preserves the global part shape while increasing local surface fidelity.

A.3 Direct Image-to-3D Generation

Direct Image-to-3D generation provides the largest geometric freedom and is mainly used for articulation-insensitive parts whose geometry is visually complex and cannot be adequately recovered through constrained refinement.

We use Hunyuan3D 2.1 (Team, 2025) to reconstruct a 3D part directly from its photorealistic part-level reference. Because direct image-to-3D generation does not explicitly preserve the original scale and coordinate frame, the generated mesh is subsequently registered to the coarse part $\mathcal{S}_i^0$. We first normalize the generated mesh by its bounding box and recover its target scale using the dimensions of the original coarse part. Its principal orientation is initialized by aligning the principal axes of the two shapes, followed by rigid Iterative Closest Point (ICP) alignment to minimize residual geometric misalignment. The aligned mesh is finally transformed back to the original part coordinate system using the recorded part transformation. We reject the generated result and request regeneration if its bounding-box aspect ratio deviates excessively from that of the original part or if the resulting alignment introduces substantial overlap with neighboring articulation-critical components.

A.4 PBR Texture Generation

PBR Texture Generation is used when the part geometry is already satisfactory but its surface appearance requires substantial improvement. We use Hunyuan3D-Paint 2.1 (Team, 2025) to synthesize physically based materials conditioned on the part-level visual reference.

Before texture generation, the input mesh is cleaned and assigned a non-overlapping UV atlas if valid UV coordinates are not already available. We use a texture resolution of 2048×2048 for all experiments. The generated material contains the base-color, roughness, and metallic channels, while geometry remains fixed throughout the operation. After generation, the texture maps are baked to the final UV atlas and evaluated by the critic for semantic correctness, cross-view consistency, and compatibility with the global object appearance. If strong seams or locally inconsistent materials are detected, the texture generation step is repeated while keeping the geometry unchanged.

A.5 Texture Projection

Texture Projection is used for parts whose surfaces are relatively flat or weakly curved and contain highly specific image-space patterns that are difficult to reproduce reliably using generative texture synthesis. Typical examples include logos, text, printed graphics, control panels, decorative paintings, and structured surface motifs.

Since the camera configuration used to generate the reference is known from the observation-planning stage, we directly project pixels from the part-level reference onto the corresponding visible mesh surface. For each pixel, a ray is cast from the reference camera to determine the first visible surface intersection. Only front-facing and unoccluded surface samples are retained, and pixels whose viewing angle with the surface normal exceeds 75° are discarded to avoid severe projection distortion. The valid projected samples are then baked into a 2048×2048 UV texture map. When multiple reference views are available, samples from different views are blended using weights based on viewing angle and visibility, giving higher weight to near-frontal observations. Unobserved texels are filled using neighboring valid colors followed by texture-space inpainting. For regions with no reliable projected evidence, we retain the original material or fall back to PBR Texture Generation.

A.6 Mesh Maintenance

Mesh Maintenance provides lightweight post-processing operations to improve the geometric validity, compactness, and downstream usability of generated parts before articulated-object reassembly. *Watertight Repair* is applied when a generated mesh contains holes, disconnected components, duplicated faces, self-intersections, or non-manifold regions. The operation first removes invalid or redundant mesh elements, repairs small boundary defects, and reconstructs problematic local regions when necessary to obtain a valid watertight surface while preserving the overall part shape. *Mesh Simplification* is applied when the generated geometry contains unnecessarily dense tessellation relative to its geometric complexity. We reduce redundant faces while preferentially preserving object boundaries, high-curvature regions, and visually important surface structures, thereby lowering mesh complexity without noticeably degrading the represented geometry.

B HUMAN EVALUATION PROTOCOL

We conduct a human study to complement the automatic evaluation with perceptual judgments of the refined articulated assets. Participants assess each result from two complementary perspectives: **Realism** and **Consistency**, both using a five-point scale. The detailed scoring criteria are defined as follows:

Realism. This criterion evaluates the overall visual quality and realism of the refined articulated asset, considering geometric quality, fine-grained details, texture and material plausibility, visual artifacts, and object-level appearance coherence.

- **1 – Very Poor:** The asset contains severe geometric or appearance errors, such as broken or highly implausible geometry, severely distorted components, or clearly incorrect textures and materials, resulting in an unrealistic or unusable appearance.
- **2 – Poor:** The asset contains substantial visual defects. The geometry lacks plausibility or important details, while textures or materials exhibit obvious artifacts, inconsistencies, or unrealistic appearance.
- **3 – Fair:** The asset is generally recognizable and visually plausible, but noticeable issues remain in geometry, surface details, textures, materials, or overall appearance coherence.
- **4 – Good:** The asset exhibits realistic geometry and appearance with rich details and coherent textures and materials, with only minor visual imperfections.
- **5 – Excellent:** The asset is highly realistic and visually coherent, with high-quality geometry, rich fine-grained details, plausible textures and materials, and no obvious visual artifacts.

Consistency. This criterion evaluates how well the refined asset preserves the original part structure and articulation of the input asset, including part identities, part composition, relative spatial relationships, geometric proportions, and articulation behavior.

- **1 – Very Poor:** The refined asset substantially deviates from the input structure. Major parts are missing, incorrectly generated, severely misplaced, or incompatible with the original articulation, resulting in obvious structural or motion failure.
- **2 – Poor:** Only a limited portion of the original structure and articulation is preserved, with substantial deviations in part geometry, scale, placement, or motion behavior.
- **3 – Fair:** The overall part composition and articulation are generally preserved, but several noticeable deviations remain in geometry, relative proportions, spatial relationships, or articulated motion.
- **4 – Good:** The refined asset largely preserves the original part structure and articulation, with only minor deviations in geometry, proportions, alignment, or motion behavior.
- **5 – Excellent:** The refined asset faithfully preserves the original part identities, composition, spatial relationships, proportions, and articulation behavior, with no noticeable structural or motion inconsistency.

Evaluation Setup. We recruit 40 participants for the human study, including 20 participants with experience in computer vision, computer graphics, or 3D content creation and 20 participants without relevant domain expertise. Each participant evaluates a randomly sampled subset of the test cases, and each result is independently rated by multiple participants.

For the **Realism** evaluation, participants are shown multi-view renderings of each refined asset and asked to assess its overall visual realism according to the criteria above. For the **Consistency** evaluation, participants are additionally shown the corresponding input articulated asset together with matched renderings and articulation animations of the refined result, allowing them to compare the part structure, relative configuration, and motion behavior before and after refinement.

To reduce potential bias, the identities of the compared methods are hidden from the participants, and the presentation order of both samples and methods is independently randomized for each participant. The same evaluation protocol is applied to all compared methods. Each participant is

Table 4: Process-level execution statistics of ArtiReal.

Statistic	Average per Asset
Observation-planning rounds	3.18
Global-reference generations	3.67
Part-reference generations	4.89
Critic evaluations	18.91
Critic rejections	8.27
Rollback operations	2.26
Retrospective-adjudication invocations	2.19
3D toolbox operations	11.37

Table 5: Usage statistics of the operations in the 3D Generation Toolbox. Here, the usage frequency denotes the percentage of refined parts for which each operation is invoked; multiple operations may be applied to the same part.

Operation	Usage Frequency
Voxel-Constrained Refinement	86.1%
Normal-Guided Refinement	32.5%
Direct Image-to-3D	24.8%
PBR Texture Generation	87.3%
Texture Projection	17.9%
Watertight Repair	34.7%
Mesh Simplification	100%

instructed to evaluate the samples independently without discussing their ratings with other participants. The final Realism and Consistency scores are computed by averaging the ratings across all participants and evaluated cases.

All participants are fairly compensated for their time at a standard hourly rate.

C Articulation Preservation Visualization

To better visualize the articulated results, we render motion videos for the articulated assets presented in the main experiments. These videos show the corresponding parts moving through their original articulation ranges and provide a more intuitive view of the refined assets under motion. The complete visualizations are provided in the supplementary material and on our anonymous supplementary page.

D System Execution Statistics

To better understand the behavior of ArtiReal during execution, we record additional process-level statistics that are not reported in the main paper. These statistics characterize how frequently the agent performs replanning, regeneration, critic-based correction, and tool invocation throughout the refinement process.

Table 4 summarizes the average execution statistics over the evaluation set. In addition to the initial forward generation process, the agent may revisit earlier stages when the critic identifies an intermediate failure. We therefore report the average numbers of observation-planning rounds, visual-reference generations, critic rejections and retrospective-adjudication invocations per asset.

We further record how frequently different operations in the 3D Generation Toolbox are selected during refinement. As shown in Tab. 5, Voxel-Constrained Refinement and PBR Texture Generation are used most frequently, while the remaining operations are selected more selectively according to the geometric and appearance characteristics of individual parts.

E LIMITATIONS

ArtiReal still has several limitations. First, the framework assumes that the input asset already provides a valid part decomposition and kinematic structure, and does not explicitly correct structural or articulation errors inherited from the source asset. Second, the final quality depends on the capabilities of the underlying image and 3D generation models, so challenging geometry, severe occlusion, or unusual materials may still lead to imperfect refinement. We will continue to explore more robust refinement strategies and stronger generation backbones to further improve the generality and reliability of ArtiReal.