PHYSAG3D: BRIDGING 3D GENERATIVE AND AGENTIC MODELS FOR HIGH-FIDELITY AND PHYSICALLY-INTERACTIVE ASSETS

Anonymous authors
Paper under double-blind review

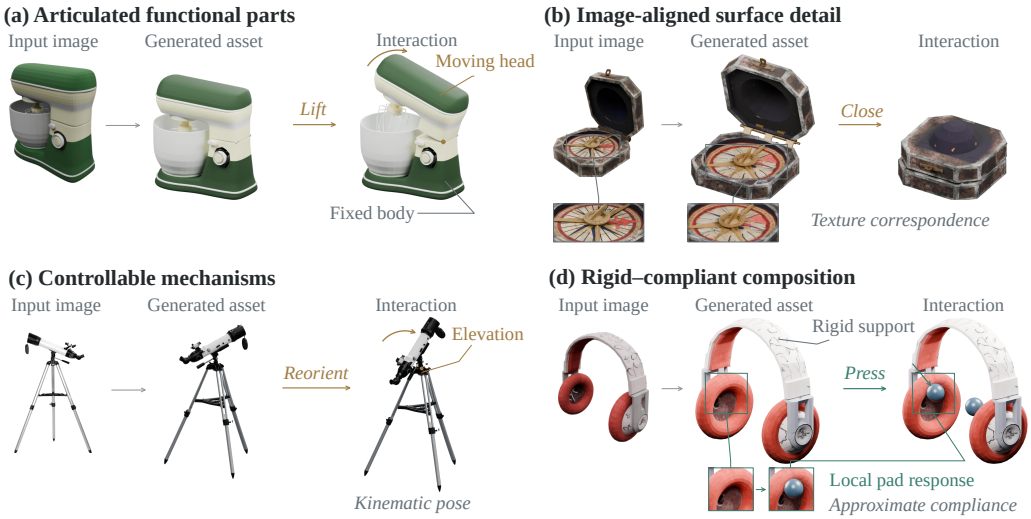

Figure 1: **Detailed appearance with region-specific interaction.** Each example shows an input image, its generated asset, and a selected interaction state. (a) A mixer head lifts relative to a fixed body. (b) A compass lid closes while retaining detailed surface appearance; insets compare dial textures. (c) Telescope joints control viewing direction. (d) Headphones combine a rigid support with locally compliant ear pads. These development examples show recorded simulated states in (a), (b), and (d), and a kinematic pose in (c); ear-pad deformation uses an approximate compliance model.

ABSTRACT

Generating interactive 3D assets from a single image requires realistic appearance, fidelity to the depicted object, and plausible physical behavior. Yet detailed geometry alone does not determine how an object should move or deform. We introduce PhysAg3D, which couples 3D generative models with agentic reasoning and tool use. A vision-language agent first infers semantic parts, candidate interactions, and physical priors from the image; a 3D generator then produces a textured whole-object mesh from the same image. A hierarchical interaction representation links these revisable priors to mechanisms and the generated surfaces they control, organizing interaction construction and refinement. Guided by this representation and observations of the mesh, the agent uses geometric and physical tools to construct parts, joints, deformable regions, and surface bindings while retaining useful generated geometry and materials. Rendered interactions and execution measurements guide targeted revisions within a fixed budget. The framework accommodates both rigid articulation and compliant deformation. Quantitative and qualitative evaluations on real photographs and external benchmarks demonstrate detailed, image-aligned assets with plausible interactions. Ablations support bounded refinement and source-surface preservation.

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

Generating interactive 3D assets from a single image offers a scalable way to populate virtual environments and embodied simulations with objects from the real world (Mildenhall et al., 2020; Kerbl et al., 2023; Mo et al., 2019; Xia et al., 2020; Xiang et al., 2020). A useful asset should capture the depicted instance, retain realistic geometric and material detail, and respond to actions in ways appropriate to the object. These goals concern different aspects of quality: a visually convincing reconstruction can remain static, while a functional mechanism can belong to an asset that no longer resembles its reference. The task therefore requires bringing appearance, instance alignment, and physical interaction into a coherent realization.

Recent advances provide strong foundations for this task. Image-conditioned generators synthesize detailed geometry and materials (Poole et al., 2023; Jun & Nichol, 2023; Nichol et al., 2022; Lin et al., 2023; Qian et al., 2023; Liu et al., 2023a; Shi et al., 2023; Liu et al., 2023b; Hong et al., 2024; Long et al., 2024; Tochil et al., 2024; Xu et al., 2024; Voleti et al., 2024; Tang et al., 2024; Xiang et al., 2025b;a); articulated and physical generation methods infer part structure, motion, and material behavior (Mo et al., 2019; Xia et al., 2020; Yu et al., 2023; Wang et al., 2019; Xiang et al., 2020; Liu et al., 2025; 2026; Cao et al., 2026; Makoviychuk et al., 2021; Hu et al., 2020); and agentic systems construct and revise assets through executable programs, visual grounding, and feedback (Ahn et al., 2022; Li et al., 2023a; Huang et al., 2023b;a; Driess et al., 2023; Brohan et al., 2023; Singh et al., 2023; Valmeekam et al., 2023; Le et al., 2025; Zhou et al., 2026; Wang et al., 2026). A remaining challenge is to maintain agreement between what an object is expected to do, how that behavior is implemented, and which visible surfaces actually respond. A single image underdetermines hidden mechanisms and physical parameters, and a generated mesh need not expose the semantic parts required for interaction. Furthermore, a valid joint or deformable simulation does not ensure that the corresponding detailed surface moves with it. Addressing these discrepancies can require changes to geometry and physical structure, creating a further tension between functional construction and preservation of the depicted appearance.

We introduce **PhysAg3D**, a framework for single-image interactive asset generation through *generative–agentic coupling*. Our central idea is to combine the instance-specific geometry and appearance supplied by a pretrained 3D generator with an agent's ability to reason about object behavior and construct executable physical models. The generated asset provides a detailed visual foundation; agentic reasoning turns this foundation into an interactive realization and revises it in response to observed behavior. This division of capabilities allows physical construction to make use of rich generated surfaces while remaining responsive to the structure and interactions of the depicted object. Figure 1 illustrates this goal through mixer-head articulation, compass-lid closure, telescope reorientation, and local ear-pad deformation. The examples associate actions with specific moving or deforming regions while retaining recognizable object structure and surface detail.

Specifically, a vision-language agent first infers revisable semantic and physical priors from the image, and a 3D generator produces a textured whole-object mesh from the same image. We propose a *hierarchical interaction representation* for this coupling, connecting semantic requirements, physical mechanisms, and concrete asset realizations. Explicit correspondences associate intended interactions with the structures and visible surfaces that implement them. Guided by these relationships and observations of the mesh, the agent uses geometric and physical tools to construct parts, joints or compliant regions, and surface bindings while retaining useful generated geometry and materials. Rendered interactions and execution measurements then guide targeted revisions with bounded construction and validation attempts. The representation expresses both rigid articulation and compliant deformation through their respective relationships between physical state and visible appearance. Our contributions are threefold:

- A generative–agentic framework that grounds interaction construction in image-derived priors while retaining useful surfaces and materials from detailed whole-object generation.
- A hierarchical interaction representation that connects requirements, mechanisms, and concrete surface realizations to organize construction and feedback-guided refinement.
- Quantitative and qualitative evaluations on real photographs and external benchmarks, assessing appearance, image alignment, and interaction attributes, with ablations of bounded refinement and source-surface preservation.

2 Related Work

Image-conditioned 3D asset generation. Early text-to-3D systems distilled diffusion priors into 3D representations (Poole et al., 2023; Jun & Nichol, 2023; Nichol et al., 2022; Lin et al., 2023; Qian et al., 2023; Tang et al., 2023); feed-forward and multi-view methods improved single-image consistency and speed (Liu et al., 2023a; Shi et al., 2023; Liu et al., 2023b; Hong et al., 2024; Long et al., 2024; Tochil et al., 2024; Xu et al., 2024; Voleti et al., 2024; Tang et al., 2024). TRELLIS introduces structured 3D latents, and TRELLIS.2 supports detailed geometry and physically based materials (Xiang et al., 2025b;a). SINGAPO conditions articulated parts on images and part relationships (Liu et al., 2025); PAct jointly generates geometry, composition, and articulation (Liu et al., 2026); PhysX-Omni covers rigid, deformable, and articulated objects (Cao et al., 2026). Shape, surface, and texture reconstruction further improve appearance (Gao et al., 2022; Li et al., 2023b; Huang et al., 2023c; Cao et al., 2023; Chen et al., 2024; Guo et al., 2020; Müller et al., 2022; Laine et al., 2020). PhysAg3D studies agent-driven construction and revision around a detailed whole-object reconstruction, linking image-derived requirements to mechanisms and visible surfaces.

Articulation and physical modeling of 3D assets. PartNet, PartNet-Mobility, GAPartNet, and Shape2Motion support research on parts, affordances, and mobility (Mo et al., 2019; Xia et al., 2020; Yu et al., 2023; Wang et al., 2019). Articulate AnyMesh segments movable parts and refines joints for open-vocabulary meshes (Qiu et al., 2025). Particulate predicts segmentation, kinematics, and motion constraints in one pass (Li et al., 2025); Instruct-Particulate adds semantic and point-prompt conditioning, combining 3D generation with vision-language inference for image-to-asset articulation (Li et al., 2026b). PhysAg3D focuses on connecting intended interactions, physical entities, exported bindings, and observed surface responses, including compliant regions whose simulation and visible geometry differ.

Agentic 3D modeling and refinement. Articulate-Anything uses vision-language reasoning and actor–critic refinement for articulated modeling (Le et al., 2025). Language-guided robotics connects instructions to executable actions and closed-loop spatial reasoning (Ahn et al., 2022; Li et al., 2023a; Huang et al., 2023b;a; Driess et al., 2023; Brohan et al., 2023), while planning studies emphasize replanning and verification (Singh et al., 2023; Valmeekam et al., 2023). ArtiCraft provides a part-and-joint SDK, object-specific tests, and structured execution feedback (Zhou et al., 2026). aDSL unifies hierarchical modeling and spatial relations for planning, execution, and critique, combining procedural scaffolds with TRELLIS through SpaceControl (Wang et al., 2026). Shape programs and CAD languages offer related compositional abstractions and persistent references (Jones et al., 2020; 2025; Cascaval et al., 2023). PhysAg3D links requirements, mechanisms, and concrete realizations to generated surfaces. Visual and execution evidence guides local revision while preserving useful geometry.

Simulation and visual grounding substrates. MuJoCo, PyBullet, Isaac Gym, and differentiable simulators provide complementary dynamics and contact substrates (Todorov et al., 2012; Coumans & Bai, 2016; Makoviychuk et al., 2021; Hu et al., 2020). Segment Anything, DINOv2, and CLIP provide visual grouping and image–text alignment primitives (Kirillov et al., 2023; Oquab et al., 2024; Radford et al., 2021). These form contextual foundations for PhysAg3D's explicit interaction-to-surface representation.

3 Method

3.1 Task Definition and Framework Overview

Given a single image I, PhysAg3D generates a textured 3D asset with object-specific interactions. An asset contains visible geometry and materials, physical entities and parameters, controls, and bindings from physical state to visible surfaces. We evaluate appearance, image alignment, and physical behavior separately; construction targets the declared interactions, since hidden structure, scale, and material parameters are not determined by one image.

Figure 2 summarizes the workflow. Image interpretation produces revisable semantic and physical hypotheses, while a pretrained generator provides a detailed whole-object mesh. Mechanism rea-

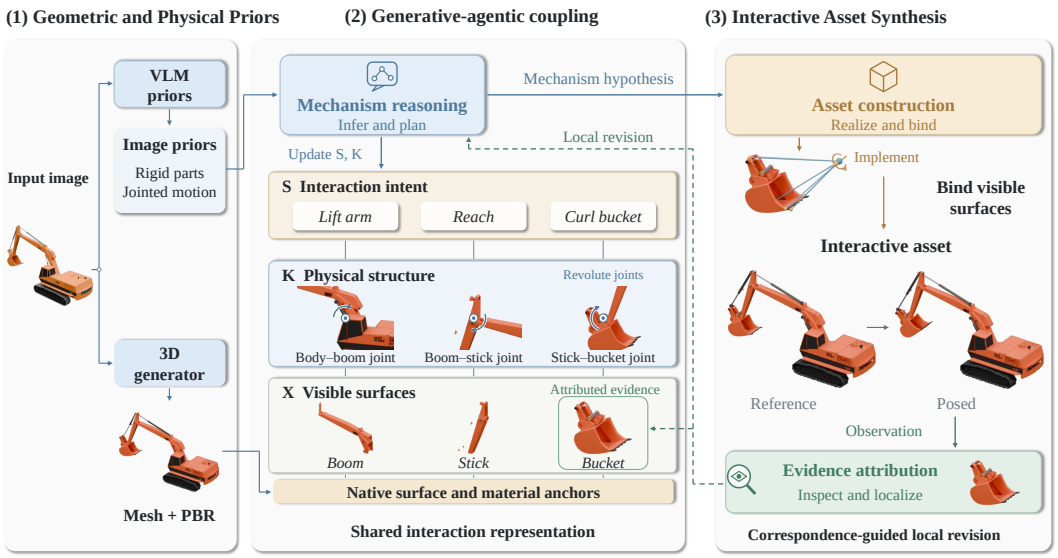

Figure 2: **Overview of PhysAg3D with an excavator example.** (1) Image-derived priors are combined with a generated textured mesh. (2) Interaction intent S specifies lifting, reaching, and curling; physical structure K specifies the corresponding joints; concrete realization X implements them with asset entities and bindings, illustrated here by its visible surfaces. (3) Asset construction binds surfaces to physical entities, and evidence attribution links observations back to the representation for local revision. Mechanism reasoning, asset construction, and evidence attribution operate within one agentic loop. Close-ups show body–boom, boom–stick, and stick–bucket connections; reference and posed states illustrate articulation.

soning connects these sources through interaction intent S (what should happen), physical structure K (which mechanism produces it), and concrete realization X (the implementing asset entities and bindings). Asset construction realizes these mechanisms and bindings; evidence attribution links observations to the responsible regions for local revision. These are functional roles in one agentic loop. Appendix A details the interaction representation, construction procedures, validation checks, and candidate-selection policy.

3.2 Generative–Agentic Coupling

Image priors and generated geometry. A vision-language agent infers provisional hypotheses $\Pi_0 = \Phi(I)$ about parts, materials, connections, candidate interactions, and scale, while retaining ambiguities. A pretrained generator produces a textured whole-object mesh $M_0 = G(I)$; our implementation uses TRELLIS.2 (Xiang et al., 2025a). The agent inspects rendered views and geometry of M_0 before fixing partitions or mechanisms. Priors guide construction but do not condition G, so the coupling occurs through interpretation, construction, and feedback without joint training.

Hierarchical interaction representation. For candidate version v, we define

$$\mathcal{R}_v = (S_v, K_v, X_v, \Gamma_v, \rho_v), \qquad \mathcal{H}_v = (\mathcal{R}_v, E_v), \tag{1}$$

where S, K, and X encode semantic requirements, mechanisms, and concrete realizations; Γ links their entities; ρ records correspondence to visible regions of M_0; and E stores candidate observations. The revision index v is distinct from execution time t.

A requirement is

$$r = (U_r, a_r, \mathcal{O}_r, \psi_r, \mathcal{D}_r), \tag{2}$$

where U_r is the target region, a_r the action protocol, $\mathcal{O}_r$ the observation, ψ_r the response test, and $\mathcal{D}_r$ the initial conditions and scope. Thus, lid opening is measured relative to its support, while local compliance is measured after removing common rigid motion. Material compliance and articulation are separate attributes.

The mechanism level K specifies rigid or compliant domains, supports, connections, degrees of freedom, joint frames, limits, and material assumptions. The realization level X identifies actual meshes, materials, scene nodes, controls, simulation entities, and exported bindings, while retaining construction or loading failures. An intended mechanism can therefore remain valid even when its realization is incomplete. In Figure 2, "Interaction intent" denotes S, "Physical structure" denotes K, and "Visible surfaces" depicts the surface component of X, whose full scope includes the implementing entities and bindings.

Correspondences across levels and surfaces. The relations Γ connect requirements to mechanisms, mechanisms to realized entities, semantic parts to visible regions, and physical drivers to bindings. The source map ρ records retained, modified, added, deleted, and unresolved surface/material regions relative to M_0. In Figure 2, the inter-level links illustrate Γ, the native surface and material anchors illustrate ρ, and attributed observations contribute to E. These maps distinguish intended relationships from exported ones: an intended hinged panel may be found attached to the base. Topology changes require an updated correspondence rather than reuse of obsolete face indices. The hierarchy expresses progressively stronger commitments about one candidate, not a fixed sequence of agent roles; requirements can select mechanisms, and observations can retrieve the associated entities for inspection.

3.3 Physical Construction with Appearance Preservation

The agent grounds semantic parts and compliant domains in image evidence, rendered views, and measurements of M_0. Connected geometry can contain multiple moving parts, and texture boundaries need not be mechanical boundaries, so every actionable surface selection is recorded in a declared frame. Construction uses Blender for geometry/material edits and physical assembly tools for simulation entities. Rigid construction sets body ownership, parent–child connections, joint frames and limits, collision geometry, and controls. Compliant construction sets a physical discretization, material assumptions, supports, loads, and a map to the visible skin. Both preserve useful source geometry, UVs, and materials while recording the purpose and source correspondence of bounded edits.

For an executable realization X,

$$z_{t+1} = F_X(z_t, u_t), \qquad V_t = B_X(z_t; V_{\text{rest}}), \qquad y_{r,t} = \mathcal{O}_{r,\Gamma,\rho}(V_t, z_t). \tag{3}$$

Here F_X is the implemented transition, B_X the exported binding, and $y_{r,t}$ the observation used to assess r. A kinematic playback and a physics simulator may produce similar visible states, but only the latter supplies dynamics evidence. Rigid vertices follow their assigned body transform; for a compliant region, one valid binding is

$$V_t = V_{\text{rest}} + W(P_t - P_{\text{rest}}), \tag{4}$$

where P_t are physical-node positions and W maps their displacements to visible vertices in a shared frame. Binding checks verify node identities, dimensions, weights, and rest geometry against the export. A correct rest pose or moving proxy alone does not establish the intended surface response.

3.4 Refinement with Visual and Physical Feedback

Each candidate is evaluated with complementary evidence: renders assess image alignment, retained appearance, part separation, and visible response; source checks assess preserved surfaces and materials; execution checks assess actions, ranges, contacts, supports, and applicable deformation or recovery. MuJoCo (Todorov et al., 2012) is used for supported physical models. A moving proxy, a joint reaching a target, and the intended visible surface responding are distinct observations. Each E_v record stores the requirement, target entities, candidate identity, action, initial state, runtime, and assessment configuration, with outcomes marked unassessed, pass, fail, or inconclusive. Errors without trajectories provide no measured response, and earlier dynamic evidence is invalidated after a build change.

For an unmet requirement, feedback combines the discrepancy or missing evidence with the related mechanisms, realized entities, and bindings obtained through Γ_v. Given remaining allowances b_v,

Table 1: Single-image interactive generation on 50 in-the-wild objects with 110 predefined actions. Image metrics use camera-grid-aligned views. ACR, MMR, and RAR are object-level macro averages; FVSP requires all actions to complete without video-visible instability. Art./Def.: articulated/deformable support; ✓/×: reported or evaluated/no reported support. Best and second-best values are bold and italic, respectively.

Method	Object Type Support		Input Consistency			Action Quality		Range	Stability
	Art.	Def.	FG-IoU↑	PSNR↑	CLIP↑	ACR↑	MMR↓	RAR↑	FVSP↑
Particulate	✓	×	*0.607*	*14.24*	*0.853*	46.7%	57.5%	34.7%	24.0%
PAct	✓	×	0.531	12.53	0.806	35.0%	87.2%	18.0%	16.0%
ArtiCraft	✓	×	0.514	12.25	0.768	56.8%	51.8%	34.3%	28.0%
PhysX-Omni	✓	✓	0.504	13.06	0.817	41.2%	68.5%	26.0%	16.0%
aDSL	✓	×	0.520	12.37	0.779	*66.8%*	*40.2%*	*48.8%*	*42.0%*
PhysAg3D (ours)	✓	✓	**0.835**	**16.47**	**0.892**	**91.3%**	**22.8%**	**91.3%**	**88.0%**

the agent proposes

$$o_v \sim \pi(I, \mathcal{R}_v, E_v, b_v), \qquad X_{v+1} = \mathrm{Apply}(X_v, o_v),$$
$$\mathcal{R}_{v+1} = \mathrm{Reconcile}(\mathcal{R}_v, o_v, X_{v+1}; M_0). \tag{5}$$

Correspondences localize inspection but do not guarantee a causal diagnosis: failures may arise from partitions, frames, supports, controls, material assumptions, bindings, or observation definitions. Reconciliation records the actual export and keeps unresolved discrepancies; accepted requirements remain traceable across revisions, so a reduced motion range cannot silently satisfy a full-closure requirement.

A declared policy bounds construction and validation and preserves prior candidates when revisions fail. The system returns the best eligible candidate under a fixed evidence-based selection rule, together with its interactive preview and unresolved requirements. A preview may be useful without a valid simulation trajectory, while its dynamics remain unverified. Appendix A gives the implementation budgets, checks, and selection policy.

4 Experiments

4.1 Experimental Settings

We evaluate single-image interactive generation on real photographs and external benchmarks, and ablate bounded refinement and source-surface preservation. Implementation and scoring details appear in Appendices A and B.

Common protocol. All comparisons fix input lists, method versions, and inference configurations; image-conditioned methods receive the same source image. We record auxiliary conditioning, output support, and computational cost. Candidate selection uses only within-run evidence, without benchmark annotations or final evaluation scores. Refinement permits at most two Builder repairs after initial construction and returns the best available candidate. All attempted inputs remain in the outcome accounting. Static assets are evaluated for input consistency; missing or invalid interaction videos count as failures within the fixed action denominator.

Qualitative development examples. Figure 1 shows separate development examples: mixer and compass articulation, approximate compliant ear pads, and telescope poses specified by joint parameters without dynamic validation.

4.2 Single-Image Interactive Generation in the Wild

Benchmark and comparisons. Our benchmark contains more than 200 real-object photographs; evaluation uses a fixed subset of 50 objects with 110 predefined semantic actions, spanning rigid, compliant, and mixed objects. Development cases are identified separately, and method-specific preprocessing is documented. We compare PhysAg3D with Particulate, PhysX-Omni, PAct, ArtiCraft, and aDSL (Li et al., 2025; Cao et al., 2026; Liu et al., 2026; Zhou et al., 2026; Wang et al., 2026). Additional qualitative comparisons appear in Appendix C.

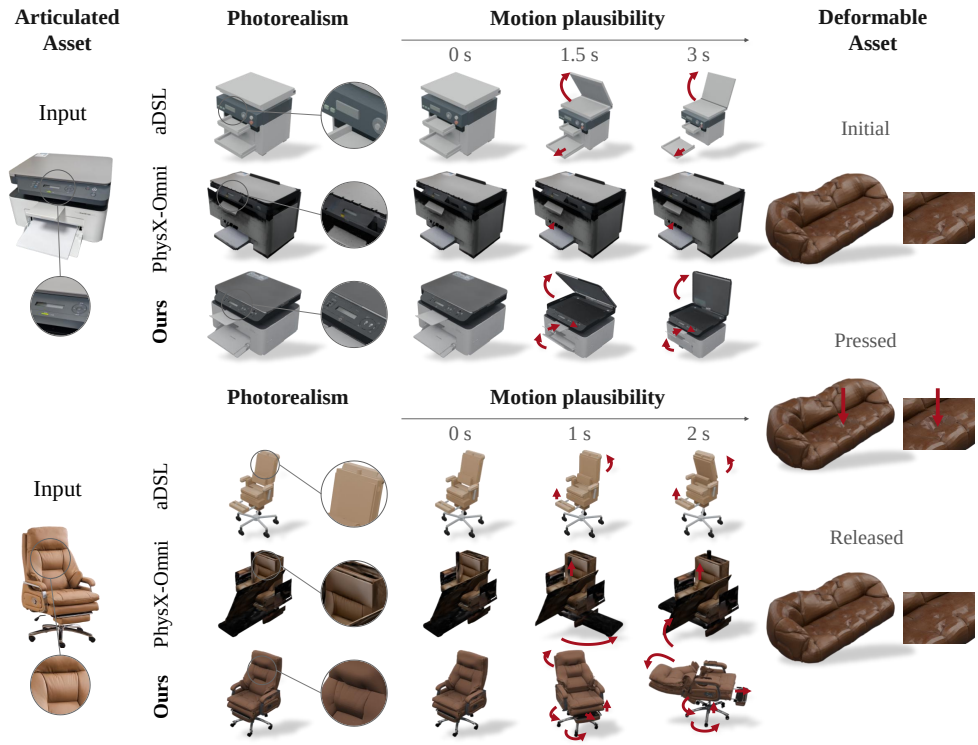

Figure 3: Qualitative comparison on real photographs. Printer and chair panels compare appearance and displayed motion across aDSL, PhysX-Omni, and PhysAg3D. The right column shows initial, pressed, and released states of a deformable asset.

Table 2: Reference-based comparison on PhysXVerse and PhysX-Mobility. CD is in 10^{-3} and F-score in 10^{-2} at threshold 0.05. Mat., Aff., Kin., and Desc. denote material, affordance, kinematics, and description. Bold and underlined values indicate the best and second-best scores. A dash indicates an entry with no reported score.

Method	PSNR↑	CD↓	F-score↑	Scale↓	Mat.↑	Aff.↑	Kin.↑	Desc.↑
Dataset: PhysXVerse								
Articulate-Anything (Le et al., 2025)	14.03	48.77	46.44	-	-	-	0.30	-
MonoArt (Li et al., 2026a)	19.68	7.03	85.27	-	-	-	0.38	-
PhysXGen (Cao et al., 2025a)	19.41	15.19	83.56	309.31	16.51	9.40	0.35	11.84
PhysX-Anything (Cao et al., 2025b)	15.97	37.06	40.46	298.19	15.65	10.50	0.42	21.38
PhysX-Omni (Cao et al., 2026)	<u>21.52</u>	<u>2.95</u>	<u>91.28</u>	<u>2.79</u>	<u>27.23</u>	21.47	0.92	<u>31.05</u>
PAct (Liu et al., 2026)	21.10	3.20	90.90	3.05	26.50	20.80	0.92	30.30
ArtiCraft (Zhou et al., 2026)	16.24	12.46	77.08	5.07	20.94	<u>27.13</u>	<u>0.94</u>	23.08
aDSL (Wang et al., 2026)	17.18	9.76	80.93	4.16	21.87	26.14	0.93	24.92
PhysAg3D (ours)	**24.26**	**2.08**	**94.34**	**2.17**	**32.18**	**28.24**	**0.96**	**38.16**
Dataset: PhysX-Mobility								
Articulate-Anything (Le et al., 2025)	15.02	16.09	66.95	-	-	-	0.64	-
MonoArt (Li et al., 2026a)	16.46	6.35	87.41	-	-	-	0.44	-
PhysXGen (Cao et al., 2025a)	15.75	35.32	79.62	46.58	16.02	8.73	0.39	11.60
PhysX-Anything (Cao et al., 2025b)	16.57	23.13	<u>89.51</u>	22.58	22.58	16.29	0.79	26.28
PhysX-Omni (Cao et al., 2026)	<u>18.38</u>	<u>4.70</u>	88.50	<u>2.78</u>	<u>24.09</u>	16.58	0.86	<u>28.40</u>
PAct (Liu et al., 2026)	18.00	5.20	86.50	3.10	23.00	16.00	0.86	27.00
ArtiCraft (Zhou et al., 2026)	15.36	10.74	78.46	5.26	17.62	<u>20.37</u>	<u>0.88</u>	22.14
aDSL (Wang et al., 2026)	16.08	8.84	81.93	4.64	19.18	19.62	0.87	23.86
PhysAg3D (ours)	**21.14**	**2.94**	**92.56**	**2.36**	**29.18**	**23.16**	**0.93**	**34.58**

Evaluation metrics. For input-view consistency, a fixed spatial grid search selects the rendered camera view closest to the input; foreground IoU, PSNR, and CLIP similarity are computed on the aligned pair. Action Completion Rate (ACR) measures correct semantic completion. Strict Mis-Motion Rate (MMR) counts missing, incomplete, incorrectly related, structurally broken, or spuriously coupled actions. Video-based Range Achievement Rate (RAR) scores the maximum valid visible range in $\{0, 0.25, 0.5, 0.75, 1\}$ after the ACR semantic gate. These three metrics are macro-averaged over objects. Function-and-Visible-Stability Pass (FVSP) is the fraction of objects

Table 3: PhysX-Bench comparison. RQS: visual quality; MCS: 3D consistency; DQS: absolute scale; APS, KPS, MPS, and DCS denote affordance, kinematic, material, and description scores. All columns are higher-is-better. Bold and underlined values indicate the best and second-best scores, including ties. A dash or NR denotes an unreported score.

Method	CLIP	RQS	MCS	DQS	APS	KPS	MPS	DCS
Articulate-Anything (Le et al., 2025)	0.55	88.46	55.27	-	-	71.25	-	-
MonoArt (Li et al., 2026a)	**0.84**	**96.20**	**82.56**	-	-	68.32	-	-
PhysXGen (Cao et al., 2025a)	<u>0.80</u>	85.93	73.50	24.21	66.07	69.17	NR	22.24
PhysX-Anything (Cao et al., 2025b)	0.55	70.81	52.71	50.20	59.96	65.99	44.70	26.89
PhysX-Omni (Cao et al., 2026)	0.77	90.00	64.48	<u>64.26</u>	70.57	80.72	59.89	<u>39.02</u>
PAct (Liu et al., 2026)	0.74	88.50	62.00	61.50	68.00	82.00	58.50	36.50
ArtiCraft (Zhou et al., 2026)	0.49	64.18	46.12	56.24	74.16	<u>85.88</u>	66.24	24.18
aDSL (Wang et al., 2026)	0.53	68.24	50.16	60.08	<u>74.82</u>	84.13	<u>67.18</u>	26.86
PhysAg3D (ours)	**0.84**	<u>95.18</u>	<u>75.24</u>	**72.16**	**82.34**	**88.18**	**72.26**	**48.12**

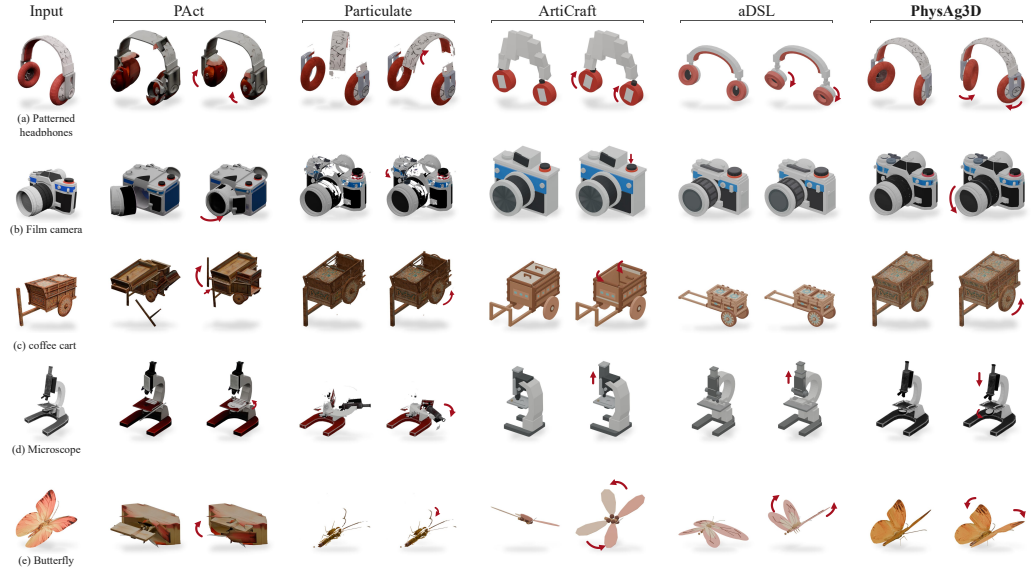

Figure 4: **Qualitative comparison on PhysXVerse.** Each row shows the input and two states per method, with fixed viewpoint and scale within each pair. Red arrows and arcs indicate translation or rotation in our results. The states show appearance and articulation playback, not calibrated dynamics or contact stability.

completing all expected actions without video-visible instability. It assesses rendered video evidence, not simulator-log stability.

PhysAg3D leads all reported input-consistency and interaction metrics in Table 1, reaching 91.3% ACR and 88.0% FVSP versus 66.8% and 42.0% for aDSL, the strongest baseline on these two measures. Figure 3 compares appearance and visible interaction states under consistent rendering conditions.

4.3 Evaluation on External Benchmarks

Reference-based evaluation. PhysXVerse and PhysX-Mobility provide reference assets for appearance, geometry, and attribute evaluation, with the dataset split reported by PhysX-Omni (Cao et al., 2026). In Table 2, PhysAg3D achieves the best reported score in every column on both datasets, improving PSNR over the strongest baseline by 2.74 and 2.76 dB, respectively.

Reference-free evaluation. PhysX-Bench evaluates generated assets using image and behavior evidence (Cao et al., 2026). We use its released evidence-generation and scoring pipeline with fixed judge configuration and prompts. In Table 3, PhysAg3D ties MonoArt on CLIP, ranks second on RQS and MCS, and leads DQS, APS, KPS, MPS, and DCS. These results show strong attribute scores alongside competitive visual quality. The articulation score uses joint playback and does not establish dynamic contact stability.

Table 4: Human perceptual evaluation over 102 valid evaluations. "#1" denotes first-place selections. "Collapse" counts evaluations in which visible motion collapse was flagged; multiple methods may be selected. Best and second-best results are bolded and underlined, respectively.

	Input-image consistency		Motion plausibility		
Method	Mean↑	#1↑	Mean↑	#1↑	Collapse↓
Particulate	<u>2.98</u>	<u>7</u>	<u>2.84</u>	<u>14</u>	21
PAct	1.99	1	1.73	0	47
ArtiCraft	2.01	1	2.06	1	28
PhysX-Omni	2.03	5	2.24	7	15
aDSL	2.38	4	2.75	11	<u>11</u>
PhysAg3D (ours)	**4.63**	**84**	**4.38**	**69**	**7**

Qualitative comparison on PhysXVerse. Figure 4 compares five selected inputs with PAct, Particulate, ArtiCraft, and aDSL (Liu et al., 2026; Li et al., 2025; Zhou et al., 2026; Wang et al., 2026). PhysAg3D preserves patterned headphones during earcup rotation, camera housing during lens extension, and ornate cart surfaces during wheel rotation; the microscope and butterfly show localized mechanism and wing motion. ArtiCraft and aDSL simplify fine geometry or patterns, while some PAct and Particulate outputs show fragmented surfaces, misplaced parts, or implausible motion.

4.4 Human Perceptual Evaluation

We evaluate six methods for input-image consistency and motion plausibility. First, second, third, and last place receive 5, 4, 3, and 1 points; unranked methods receive 2. First and last place are required, whereas second and third are optional. A separate multi-select question records visible motion collapse.

Across 102 valid evaluations (Table 4), PhysAg3D leads in mean input-consistency (4.63 versus 2.98) and motion-plausibility scores (4.38 versus 2.84), and first-place selections. Collapse is flagged 7 times versus 11–47 for baselines. These judgments measure perceived fidelity and motion plausibility, not calibrated dynamics or contact stability.

4.5 Ablation Study

Ablation settings. The three settings share PhysX-Bench inputs, cached TRELLIS.2 reconstructions, image priors, model and decoding settings, construction tools, and final scoring. *Full pipeline* preserves useful source geometry, UVs, and materials while allowing evidence-guided edits and up to two Builder repairs using visual and physical feedback. *One-shot construction* returns the same initial build without repair, but still undergoes final evaluation. *w/o surface preservation* removes the construction-time geometry, UV, and material preservation constraints while retaining the initial reconstruction, physical-to-visible bindings, repair budget, and candidate-selection rule.

Table 5: Ablation on PhysX-Bench.

Setting	CLIP↑	RQS↑	KPS↑	MPS↑
w/o surface preservation	0.68	77.50	75.00	59.00
One-shot construction	0.72	82.80	68.50	55.00
Full pipeline	**0.84**	**95.18**	**88.18**	**72.26**

Results. Table 5 shows that Full pipeline improves KPS and MPS over One-shot by 19.68 and 17.26 points. Removing surface preservation reduces CLIP by 0.16 and RQS by 17.68. These results support bounded refinement and source preservation; failed and inconclusive runs remain in the evaluation.

5 Conclusion

We presented PhysAg3D for generating interactive 3D assets from a single image. Its hierarchical representation connects interaction requirements, physical mechanisms, and visible surfaces to guide construction and feedback-based refinement. Experiments support appearance preservation and plausible interactions, with ablations indicating benefits from refinement and surface preservation. Recovering hidden structure and material properties remains a limitation.

STATEMENT

We used generative AI tools for language polishing and assistance with repeated ablation experiment runs and parameter adjustments. The authors manually inspected the execution logs, verified the experimental outputs and reported results, and take responsibility for the final content of this work.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. URL https://arxiv.org/abs/2204.01691.

Anthony Brohan, Noah Brown, Justice Carbajal, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. URL https://arxiv.org/abs/2307.15818.

Chenjie Cao, Nikolaos Sarafianos, Tony Tung, et al. TexFusion: Synthesizing 3d textures with text-guided image diffusion models. *arXiv preprint arXiv:2310.13766*, 2023. URL https://arxiv.org/abs/2310.13766.

Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. PhysX-3D: Physical-grounded 3D asset generation. *arXiv preprint arXiv:2507.12465*, 2025a. URL https://arxiv.org/abs/2507.12465.

Ziang Cao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. PhysX-Anything: Simulation-ready physical 3D assets from single image. *arXiv preprint arXiv:2511.13648*, 2025b. URL https://arxiv.org/abs/2511.13648.

Ziang Cao, Yinghao Liu, Haitian Li, Runmao Yao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. PhysX-Omni: Unified simulation-ready physical 3D generation for rigid, deformable, and articulated objects. *arXiv preprint arXiv:2605.21572*, 2026. URL https://arxiv.org/abs/2605.21572.

Dan Cascaval, Rastislav Bodik, and Adriana Schulz. A lineage-based referencing DSL for computer-aided design. *Proceedings of the ACM on Programming Languages*, 7(PLDI):109:1–109:24, 2023. doi: 10.1145/3591223. URL https://doi.org/10.1145/3591223.

Rui Chen et al. GenesisTex: Adapting image denoising diffusion to texture space. *arXiv preprint arXiv:2403.17782*, 2024. URL https://arxiv.org/abs/2403.17782.

Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. *Technical report*, 2016. URL https://pybullet.org.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, et al. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. URL https://arxiv.org/abs/2303.03378.

Jun Gao, Tianchang Shen, Zian Wang, et al. GET3D: A generative model of high quality 3D textured shapes learned from images. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2209.11163.

Yu Guo, Xin Li, Zhaowen Zhu, et al. MaterialGAN: Reflectance capture using a generative network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2011.12660.

Jun Hong, Bokui Mei, Jianyuan Zeng, et al. LRM: Large reconstruction model for single image to 3D. *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.04400.

Yuanming Hu, Luke Anderson, Tzu-Mao Li, et al. Difftaichi: Differentiable programming for physical simulation. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.00935.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2023a. URL https://arxiv.org/abs/2207.05608.

Wenlong Huang, F. Xia, T. Xiao, et al. VoxPoser: Composable 3D value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023b. URL https://arxiv.org/abs/2307.05973.

Zheng Huang, Xianming Wang, et al. Text2Tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023c. URL https://arxiv.org/abs/2303.11396.

Ben T. Jones, Zihan Zhang, Felix Hähnlein, Wojciech Matusik, Maaz Ahmad, Vladimir Kim, and Adriana Schulz. A solver-aided hierarchical language for LLM-driven CAD design. *Computer Graphics Forum*, 44(7):e70250, 2025. doi: 10.1111/cgf.70250. URL https://doi.org/10.1111/cgf.70250.

R. Kenny Jones, Theresa Barton, Xianghao Xu, Kai Wang, Ellen Jiang, Paul Guerrero, Niloy J. Mitra, and Daniel Ritchie. ShapeAssembly: Learning to generate programs for 3D shape structure synthesis. *ACM Transactions on Graphics*, 39(6):234:1–234:20, 2020. doi: 10.1145/3414685.3417812. URL https://doi.org/10.1145/3414685.3417812.

Heewoo Jun and Alex Nichol. Shap-E: Generating conditional 3D implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. URL https://arxiv.org/abs/2305.02463.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM Transactions on Graphics*, volume 42, 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. URL https://arxiv.org/abs/2304.02643.

Samuli Laine, Janne Hellsten, Tero Karras, Yeongwoo Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. In *ACM Transactions on Graphics*, volume 39, 2020. URL https://research.nvidia.com/labs/rtr/nvdiffrast/.

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. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.13882.

Chengshu Li, Jack Liang, Wenlong Shi, et al. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2023a. URL https://arxiv.org/abs/2209.07753.

Haitian Li, Haozhe Xie, Junxiang Xu, Beichen Wen, Fangzhou Hong, and Ziwei Liu. MonoArt: Progressive structural reasoning for monocular articulated 3D reconstruction. *arXiv preprint arXiv:2603.19231*, 2026a. URL https://arxiv.org/abs/2603.19231.

Ruining Li, Yuxin Yao, Chuanxia Zheng, Christian Rupprecht, Joan Lasenby, Shangzhe Wu, and Andrea Vedaldi. Particulate: Feed-forward 3D object articulation. *arXiv preprint arXiv:2512.11798*, 2025. URL https://arxiv.org/abs/2512.11798.

Ruining Li, Yuxin Yao, Matt Zhou, Chuanxia Zheng, Christian Rupprecht, Joan Lasenby, Shangzhe Wu, and Andrea Vedaldi. Instruct-Particulate: Scaling feed-forward 3D object articulation with kinematic control. *arXiv preprint arXiv:2606.14699*, 2026b. URL https://arxiv.org/abs/2606.14699.

Zhaoshuo Li, Thomas Müller, Alex Evans, et al. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b. URL https://arxiv.org/abs/2306.03092.

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*, 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 objects. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.16499.

Qingming Liu, Xinyue Yao, Shuyuan Zhang, Yueci Deng, Guiliang Liu, Zhen Liu, and Kui Jia. PAct: Part-decomposed single-view articulated object generation. *arXiv preprint arXiv:2602.14965*, 2026. URL https://arxiv.org/abs/2602.14965.

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. *arXiv preprint arXiv:2303.11328*, 2023a. URL https://arxiv.org/abs/2303.11328.

Yuan Liu, Cheng Lin, Zhong Zeng, Xiaoxiao Long, et al. SyncDreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023b. URL https://arxiv.org/abs/2309.03453.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3D: Single image to 3D using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2310.15008.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, et al. Isaac Gym: High performance GPU-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. URL https://arxiv.org/abs/2108.10470.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2003.08934.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subham Tripathi, and Leonidas J. Guibas. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1812.02713.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4), 2022. URL https://arxiv.org/abs/2201.05989.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-E: A system for generating 3D point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. URL https://arxiv.org/abs/2212.08751.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. In *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14988.

Guocheng Qian, Jinjing Mai, Hasan Hammoud, Antonio Alliegro, Cordelia Schmid, and Bernard Ghanem. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. URL https://arxiv.org/abs/2306.17843.

Xiaowen Qiu, Jincheng Yang, Yian Wang, Zhehuan Chen, Yufei Wang, Tsun-Hsuan Wang, Zhou Xian, and Chuang Gan. Articulate AnyMesh: Open-vocabulary 3D articulated objects modeling. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 4585–4606, 2025. URL https://proceedings.mlr.press/v305/qiu25b.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2103.00020.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. *arXiv preprint arXiv:2308.16512*, 2023. URL https://arxiv.org/abs/2308.16512.

Ishika Singh, Valts Blukis, Arsalan Mousavian, et al. ProgPrompt: Generating situated robot task plans with large language models. *arXiv preprint arXiv:2209.11302*, 2023. URL https://arxiv.org/abs/2209.11302.

Jiaxiang Tang et al. LGM: Large multi-view gaussian model for high-resolution 3D content creation. *arXiv preprint arXiv:2402.05054*, 2024. URL https://arxiv.org/abs/2402.05054.

Junshu Tang, Meng Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Make-It-3D: High-fidelity 3D creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. URL https://arxiv.org/abs/2303.14184.

Sergey Tochil et al. TripoSR: Fast 3D object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. URL https://arxiv.org/abs/2403.02151.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109. URL https://doi.org/10.1109/IROS.2012.6386109.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *arXiv preprint arXiv:2206.10498*, 2023. URL https://arxiv.org/abs/2206.10498.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, Albert Pumarola, Varun Jampani, and Yan Yan. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. URL https://arxiv.org/abs/2403.12008.

Rui-Huan Wang, Si-Tong Wei, Jia-Qi He, Heng-Yi Wei, Baoquan Chen, and Peng-Shuai Wang. aDSL: Agentic 3D creation via joint agent-program design. *arXiv preprint arXiv:2608.17975*, 2026. URL https://arxiv.org/abs/2608.17975.

Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2Motion: Joint analysis of motion parts and attributes from 3D shapes. *arXiv preprint arXiv:1903.03911*, 2019. URL https://arxiv.org/abs/1903.03911.

Fei Xia, Yixin Zhu, Konstantinos Bousmalis, et al. PartNet-Mobility: A benchmark for learning articulated object models from point clouds. *arXiv preprint arXiv:2004.04880*, 2020. URL https://arxiv.org/abs/2004.04880.

Fu Xiang, Frederik Ebert, Corey Cheng, et al. SAPIEN: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2003.08515.

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. *arXiv preprint arXiv:2512.14692*, 2025a. URL https://arxiv.org/abs/2512.14692.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21469–21480, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xiang_Structured_3D_Latents_for_Scalable_and_Versatile_3D_Generation_CVPR_2025_paper.html.

Jiale Xu et al. InstantMesh: Efficient 3D mesh generation from a single image. *arXiv preprint arXiv:2404.07191*, 2024. URL https://arxiv.org/abs/2404.07191.

Qiaojun Yu, Ce Hao, Xibin Yuan, et al. GAPartNet: Cross-category generalizable object perception and manipulation using 3D part affordance. *arXiv preprint arXiv:2211.05272*, 2023. URL https://arxiv.org/abs/2211.05272.

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. URL https://arxiv.org/abs/2605.15187.

SUPPLEMENTARY MATERIAL

This supplement provides implementation details, evaluation protocols, and an extended qualitative comparison accompanying the main paper.

A METHOD AND IMPLEMENTATION DETAILS

A.1 INTERACTION REPRESENTATION

The representation in Section 3.2 separates interaction intent, mechanism design, and concrete assets. The semantic level S identifies the target region, action, expected response, and observation frame. The mechanism level K specifies bodies or compliant domains, connections, degrees of freedom, supports, and parameter assumptions. The realization level X identifies the meshes, materials, scene entities, controls, and exported bindings. The relations Γ connect these levels; ρ records the origin and modification of source surfaces; and E associates observations with the candidate that produced them. These correspondences link interaction requirements to mechanisms, realized entities, and surface bindings for construction and feedback-guided revision, as defined in Section 3.2. The following subsections detail the associated construction and validation procedures.

For example, a lid-opening requirement refers to the lid region and measures its motion relative to the base. Its mechanism identifies the hinge axis and permitted range, while its realization identifies the moving body and the visible lid faces assigned to it. This correspondence allows the hinge configuration and visible-surface assignment to be inspected separately; it localizes the relevant entities without by itself establishing the cause of a failure. For a padded region, the corresponding records identify the load, support, physical discretization, and visible skin. These examples explain the representation without assuming that hidden mechanisms or material parameters are uniquely recoverable from the input image.

A.2 GEOMETRY, MATERIALS, AND SURFACE BINDING

The initial TRELLIS.2 reconstruction supplies the detailed visible geometry and materials. Geometric refinement uses rendered views and measurements to identify actionable regions: mesh connectivity alone does not determine which parts should move, and a texture boundary need not be a mechanical boundary. Bounded edits preserve useful source faces, UV coordinates, and material assignments. When topology changes, source correspondence must be updated rather than inherited from obsolete face indices.

For rigid articulation, visible vertices follow the transform of their assigned body. Correct body ownership, local transforms, parent–child relationships, joint frames, and limits are all needed to produce the intended visible motion. Collision geometry can differ from the detailed render mesh, but motion of a collision proxy alone does not verify motion of the visible part.

For compliant regions, the physical discretization and visible skin may also differ. The displacement binding in Eq. 4 transfers physical-node displacements to visible vertices in a common frame. Its node identities, dimensions, weights, and rest geometry are checked against the exported asset. Supports and loading conditions are specified separately from the skin binding. A local deformation observation removes common rigid motion before assessing shape change; otherwise, movement of the entire object could be mistaken for local compliance. The ear-pad example uses an approximate compliance model rather than a material calibration recovered from the image.

A.3 FEEDBACK AND CANDIDATE RECORDS

Rendered observations assess appearance, part separation, and visible response. Execution observations assess the requested action, achieved range, contacts, supports, and applicable deformation or recovery. Each record identifies the requirement, candidate, target region and entities, action, initial state, runtime and assessment configurations, and assessment outcome. The four outcomes are unassessed, pass, fail, and inconclusive. An execution error supplies no measured trajectory, and a changed build requires renewed dynamic evidence.

Table 6: Documented implementation configuration and construction budgets. Host validation attempts and simulations called inside construction tools are counted separately.

Component	Configuration
3D generator	TRELLIS.2-4B; `1024_cascade` route
Sampling and texture	24 sampling steps; 2048-pixel texture size
Generator seed	42
Reasoning model	`gpt-6-astra`; high reasoning effort
Geometry / simulation	Blender / MuJoCo
Builder attempts	3 total: initial construction and at most 2 repairs
Host physical validations	At most 2 attempts
Structural replanning	At most 1 replan
Planner / Reviewer responses	At most 2 / 3 responses
Returned role messages	At most 28
Joint-range reductions	At most 2 per joint

Feedback associates a discrepancy with the relevant mechanism, realized entity, or binding. Local repairs can address motion range, tracking, surface assignment, or assembly; structural replanning addresses topology, ownership, joint-axis, or material contradictions. Previous candidates remain available when a revision fails. The budgets and selection procedure below bound this loop.

A.4 Runtime Configuration and Budgets

Table 6 summarizes the documented v0.2.0 runtime configuration. Blender handles geometry and material editing; MuJoCo executes supported physical models. The reasoning model is accessed through the local Codex interface. The configured generator seed identifies the initialization setting, without implying deterministic agent execution.

Invalid returned messages consume their role-response allowance. Host validation attempts are reserved before execution, so an interrupted simulation also consumes an attempt. Candidate history and cumulative counters persist across continuations; selecting an earlier candidate does not reset them. Simulator calls made inside construction tools are not covered by the host-attempt limit. These counts therefore describe distinct resources rather than an aggregate computation budget. The initial construction and at most two repairs constitute the Builder budget in Section 4.1; structural replanning does not reset this budget. As in Section 3.4, a modified build requires new dynamic evidence, and a preview does not turn an unverified simulation status into a pass.

A.5 Motion-Range Adaptation and Physical Checks

For a hinge or slide with original interval $[\ell_0, h_0]$ and neutral coordinate q_0, the documented range helper tests

$$\ell_n = q_0 + (\ell_0 - q_0)(2/3)^n, \qquad h_n = q_0 + (h_0 - q_0)(2/3)^n, \quad n \in \{0, 1, 2\}. \tag{6}$$

The original interval remains the reference across repairs. The retained span must be at least $0.2(h_0 - \ell_0)$ and, for a hinge, at least 10°. A reduced interval is considered only after a completed failed sweep with a valid rest pose. The checks include rest and both motion directions where applicable. A colliding rest pose, missing clearance information, incomplete sweep, or execution error terminates this search. The achieved range is retained explicitly; reducing travel does not satisfy an unchanged full-closure requirement.

Let D be the positive asset diagonal in meters. The documented acceptance profile uses a contact-proxy penetration tolerance of $0.005D$, an angular endpoint tolerance of 5°, and a translational endpoint tolerance of $0.005D$. Additional checks cover finite states, solver warnings, joint limits, measured interaction, and applicable compliant-region requirements. Geometric sweep checks apply to sampled configurations and do not replace dynamic execution.

For a declared compliant region, a least-squares rigid alignment removes common motion from the simulation nodes. Let $d_{\max}$ be the maximum residual over recorded samples and phase endpoints. The response criterion requires $d_{\max} \geq 0.001D$ together with a pairwise-distance-change check.

When recovery is required, the residual after unloading must not exceed $\max(0.001D, 0.35d_{\max})$. These criteria are deformation heuristics on simulation nodes; exported-surface response is assessed separately through the binding and visible observations.

A.6 Output Selection

Selection excludes unloadable, empty, nonfinite, externally dependent assets and detected source-identity conflicts. In autonomous runs, eligible candidates are ranked by appearance evidence, native retention, semantic evidence, absence of unsupported additions, display checks, declared bodies, supported physics acceptance, and recency, in that order. Native retention includes source identity, face accounting, geometry, and UV/material checks. As specified in Section 4.5, Full pipeline and w/o surface preservation use the same selection rule. In the latter, source-geometry, UV, and material preservation constraints are removed from construction, while the shared ranking criteria and physical-to-visible bindings remain. This intervention therefore tests the construction constraints under a fixed output-selection policy. Supported physics acceptance requires a passing report, at least three trajectory states, and matching asset and trajectory identities. A preview can be returned without a valid dynamic trajectory, with that status kept distinct. If no refined candidate is eligible, the native reconstruction can serve as a static fallback. Such a fallback is not evidence that the requested interaction or dynamic validation succeeded; failed and inconclusive attempts remain in the outcome accounting.

B Evaluation and Comparison Details

B.1 Evaluation Scope and Aggregation

The in-the-wild collection contains more than 200 photographs; the fixed evaluation subset contains 50 objects and 110 predefined semantic actions. Following Section 4.2, ACR, MMR, and RAR are macro-averaged over objects. For normalized object-level scores $s_i^{(m)}$, metric m is reported as $100N^{-1}\sum_{i=1}^{N} s_i^{(m)}$, with $N = 50$. Thus, each object contributes equally, regardless of its number of rendered frames.

ACR measures correct semantic action completion. Strict MMR counts missing, incomplete, incorrectly related, structurally broken, or spuriously coupled actions. RAR scores the maximum valid visible range in $\{0, 0.25, 0.5, 0.75, 1\}$ after the ACR semantic gate. FVSP is the percentage of objects that complete every expected action without video-visible instability. It is a joint functional-and-visible-stability criterion based on rendered videos, not simulator logs.

Support marks indicate declared or evaluated capabilities, not successful outputs. Static assets remain evaluable for input consistency; missing or invalid interaction videos count as failures within the fixed action denominator. Failed and inconclusive attempts remain in the outcome accounting, with unsupported output types distinguished in the records. Development examples remain separate from the benchmark comparison.

B.2 Inputs, Baselines, and Scoring

Section 4.1 specifies the common comparison protocol. Image-conditioned methods receive the same photograph; mesh-conditioned methods require an identified reconstruction and any auxiliary conditioning. Format and coordinate conversions are distinguished from geometry repair or inference of missing physical parameters, which would change the evaluated method. PhysXGen denotes the framework introduced in PhysX-3D (Cao et al., 2025a). aDSL reports image-to-shape generation and articulated creation as a downstream application, motivating its inclusion in the single-image comparison.

Reference-based comparisons use compatible alignment, pose, sampling, distance normalization, and scoring settings. Shape normalization must preserve the separately assessed absolute-scale error. Table 2 follows the PhysXVerse/PhysX-Mobility split reported by PhysX-Omni; its kinematic score is not a dynamics success rate. Missing scores, marked by a dash or NR, are unreported entries rather than measured zeros.

PhysX-Bench uses fixed evidence generation, evaluator configuration, and prompts across methods. Joint playback supplies articulation evidence, while material response and static appearance are separate observations; playback does not establish dynamic contact stability. The selected states in Figure 4 likewise demonstrate visible articulation and surface appearance. Table 7 summarizes the evaluation settings.

B.3 Ablation Configuration

The three variants in Section 4.5 share inputs, cached reconstructions, initial priors, model and decoding settings, construction tools, and final scoring. Full pipeline and w/o surface preservation share the two-repair limit and candidate-selection rule. One-shot returns the initial candidate and receives final scoring without feeding it back for repair. All attempted inputs, including failed and inconclusive runs, remain in the outcome accounting.

The source-preservation ablation removes construction-time geometry, UV, and material constraints while retaining the initial reconstruction and physical-to-visible bindings. It keeps output selection fixed, including the native-retention ranking criterion. The One-shot comparison measures refinement as a whole, including additional repair and candidate-selection opportunities; it does not isolate an individual feedback signal.

C Additional Qualitative Comparisons

Figure 5 extends Figure 3 with additional cases and methods, showing geometry, materials, and texture continuity alongside input-image agreement in proportions, parts, colors, and patterns. No additional aggregate score is introduced.

Table 7: Evidence and interpretation for the evaluations reported in the main paper. The rows describe evaluation settings, not additional experiments.

Evaluation	Evidence and interpretation
In-the-wild interaction	ACR, MMR, and RAR: macro averages over 50 objects, computed from 110 predefined actions. FVSP: object-level pass rate requiring all actions to complete without video-visible instability.
PhysXVerse / PhysX-Mobility	Reference assets support geometry, appearance, scale, and attribute comparisons. CD uses 10^{-3} units; F-score uses 10^{-2} units at threshold 0.05, as in Table 2.
PhysX-Bench	CLIP, RQS, and MCS address alignment, visual quality, and multi-view consistency. DQS addresses scale; APS, KPS, MPS, and DCS address affordance, kinematics, materials, and description.
Component ablation	The same PhysX-Bench setting compares Full pipeline, One-shot, and w/o surface preservation using CLIP, RQS, KPS, and MPS.
Extended visualization	Additional real-world cases and methods support inspection of photorealistic appearance and correspondence to the input image.

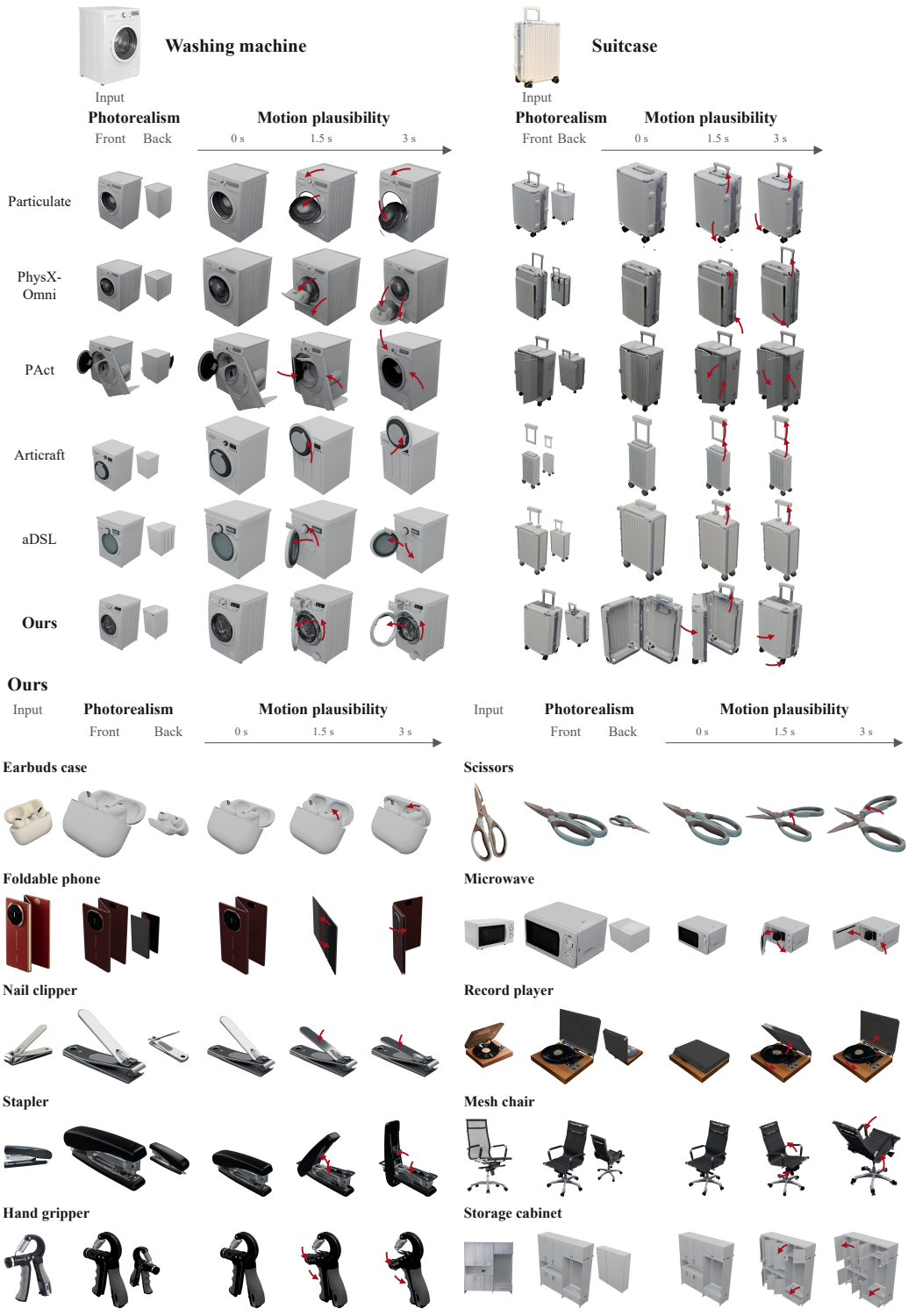

Figure 5: **Extended qualitative comparison on real-world images.** Complementing Figure 3, the expanded comparison focuses on photorealistic appearance and alignment with the input photograph across additional cases and methods.