PHI-ART: PHYSICS-AWARE ARTICULATED OBJECT GENERATION FROM A SINGLE IMAGE WITH A MULTIMODAL LARGE LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

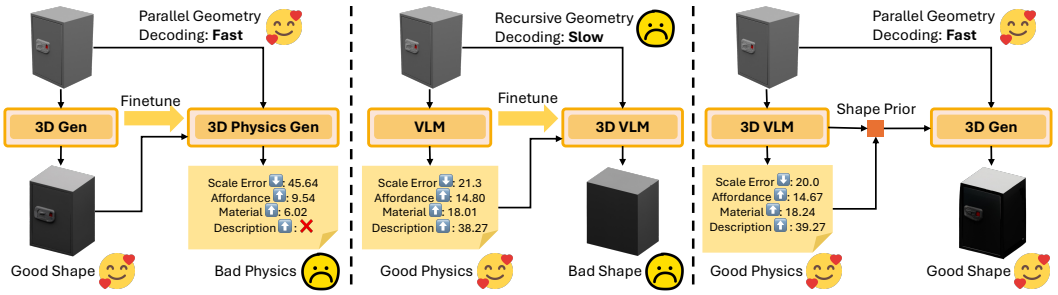

Figure 1: (a) 3D-generation-oriented methods preserve shape fidelity but fail in physical information prediction. (b) VLM-based methods predict good physical attributes but can sacrifice geometric fidelity and require recursive queries. (c) Our method uses a 3D VLM predicts physical attributes and a shape prior that guides detailed shape generation, preserving geometric detail and physical information without recursive geometry queries.

ABSTRACT

Generating articulated assets from a single image requires more than reconstructing visible surfaces. Part geometry, kinematic structure, and physical attributes must be inferred from incomplete visual evidence to support interactive simulation. However, existing models are struggling to tackle these problem all together. Thus, we introduce Phi-Art: ϕ-Art, a multimodal framework that couples structured object prediction with parallel geometry decoding. A multimodal language model predicts part semantics, 3D bounding boxes, joint parameters, and physical attributes, providing a coarse spatial plan for asset generation. A Qformer-style decoder combines the model's hidden representations with image features to predict whole-object geometry codes in parallel, separating geometric decoding from autoregressive text generation. We further corrupt structured descriptions during training to encourage the use of visual evidence and reduce reliance on accurate textual layouts. The predicted geometry and layouts guide part synthesis, followed by kinematic assembly and export to simulator formats. Experiments on PhysXNet and Mobility show that our method outperforms the existing baselines in the geometry accuracy, training and inference efficiency with a competitive physical attributes prediction. We will release our code, model checkpoints, and processed training data.

1 INTRODUCTION

Recent progress in single-image 3D generation has made it possible to recover detailed geometry and appearance from ordinary photographs (Hong et al., 2023; Xiang et al., 2024). Turning these visual assets into objects that can be manipulated in simulation requires a richer description. An articulated asset must specify its constituent parts, their spatial arrangement, the joints that govern their motion, and the physical attributes associated with each component. Recent methods have extended 3D generation to part decomposition, articulation, and physical attributes (Yang et al., 2025; Le et al., 2024; Cao et al., 2025a;b). Figure 1(a,b) illustrates two approaches to this goal. 3D-generation-oriented methods extend learned geometric representations to physical-attribute prediction. VLM-based methods instead express physical and geometric information through structured descriptions. These approaches offer complementary strengths: the geometric priors of 3D generators and the

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.

structured semantic and physical prediction of VLMs. The challenge is to combine these strengths together to enhance the physics-aware articulated object generation.

Our motivation stems from the observation that an object's structure can be described more coarsely than its geometry. A cabinet, for example, can be described by its frame, doors, and hinge relationships before the precise surface of each component is specified. Human vision provides a related motivation: global organization can take precedence over local features (Navon, 1977), while structural theories of object recognition describe shapes through components and their spatial relations (Marr & Nishihara, 1978; Biederman, 1987). For articulated generation, this suggests separating what the object is made of and how its parts are organized from what those parts actually look like. The former provides a structural plan; the latter should remain grounded in the input image.

We introduce ϕ-Art, which combines VLM-based structural and physical prediction with image-conditioned 3D generation, as illustrated in Figure 1(c). Given a single image, a multimodal language model predicts part semantics, 3D bounding boxes, joint parameters, and physical attributes. The boxes locate the predicted parts in a shared coordinate system and provide a coarse spatial plan for generation. A query-based geometry decoder combines the model's hidden representation with dense image features to predict whole-object geometry codes in parallel, separating geometric decoding from autoregressive structured prediction. The decoded occupancy provides a coarse shape prior that, together with the predicted part boxes, input image, and visible-part masks, conditions a pretrained part generator (Yang et al., 2025) to synthesize detailed part geometry and appearance.

Structural guidance can nevertheless become a shortcut, allowing the geometry decoder to rely on bounding-box context while ignoring the image. Our initial analyses reveal this behavior. To keep geometry grounded in visual evidence, we combine direct image conditioning with training-time regularization. A frozen DINOv3 encoder (Siméoni et al., 2025) extracts dense image features, which a Q-Former-style module aggregates with the language model's hidden representation for parallel geometry decoding. However, providing image features alone does not ensure that the decoder uses them. We therefore corrupt the bounding-box context during training while preserving the image and target geometry, reducing dependence on accurate textual layouts. We further impose a silhouette loss that penalizes disagreement between the projection of decoded occupancy and the foreground silhouette of the conditioning image. Together, these mechanisms encourage the decoder to use the structural plan as guidance while recovering instance-specific geometry from visual evidence. Controlled comparisons on PhysXNet and Mobility show improvements over existing baselines in both geometry and physical attributes. Our contributions are summarized as follows:

- We introduce ϕ-Art, a framework that generates assets from a single image, integrating geometry, appearance, part and physical attributes, and articulation within a single pipeline.
- We develop a Q-Former-style geometry decoder that fuses language-model representations with dense image features to predict whole-object geometry codes in parallel.
- We introduce bounding-box corruption and silhouette supervision to encourage image-grounded generation, reducing reliance on textual layouts while promoting consistency between the generated geometry and the conditioning image.
- We achieve state-of-the-art performance in articulated asset generation, outperforming existing methods in geometric accuracy and comparable in physical-attribute prediction.

2 Related Work

2.1 3D Object Generation

Native 3D generation models geometry and appearance through latent diffusion, flow matching, or autoregressive prediction. 3DShape2VecSet (Zhang et al., 2023) introduces vector sets for neural fields, while Michelangelo (Zhao et al., 2023) aligns shape representations with images and text. CLAY (Zhang et al., 2024) scales latent diffusion with a multi-resolution VAE; TRELLIS (Xiang et al., 2024) generates structured sparse latents. TripoSG (Li et al., 2025b) scales SDF-based rectified flow, and Hunyuan3D 2.0 (Zhao et al., 2025) separates shape and texture synthesis. MeshGPT (Siddiqui et al., 2024) autoregressively generates mesh tokens, while ShapeLLM-Omni (Ye et al., 2025) unifies text, images, and 3D tokens. These methods emphasize holistic geometry and appearance over functional structure. Building on ShapeLLM-Omni, ϕ-Art further predicts part layouts, kinematics, and physical attributes, using the resulting context to guide parallel coarse geometry decoding.

Table 1: **Capability comparison with representative methods.** Checks ✓ and crosses × indicate supported and unsupported capabilities in the compared core pipelines, respectively.

Capability	ShapeLLM-Omni	OmniPart	URDFormer	Articulate-Anything	URDF-Anything+	PAct	PhysX-3D	PhysX-Anything	**ϕ-Art (ours)**
Photorealistic	✓	✓	×	×	×	✓	✓	✓	✓
Shape generation	✓	✓	×	×	✓	✓	✓	✓	✓
Part modeling	×	✓	✓	✓	✓	✓	×	✓	✓
Articulation	×	×	✓	✓	✓	✓	×	✓	✓
Physical attributes	×	×	×	×	×	×	✓	✓	✓
Spatial layout	×	✓	×	×	×	×	×	×	✓
Part grounding	×	×	×	×	×	×	×	×	✓

2.2 Part-Aware and Articulated 3D Modeling

Part-aware modeling enables generation beyond monolithic geometry. PartNet (Mo et al., 2019) provides hierarchical part annotations, and SAPIEN (Xiang et al., 2020) introduces articulated PartNet-Mobility assets and simulation support. PartCrafter (Lin et al., 2025) jointly generates part meshes for a specified part count using compositional latents, while PartPacker (Tang et al., 2025) represents parts through dual volume packing. OmniPart (Yang et al., 2025) combines part-box planning with joint geometry and appearance synthesis. Articulate-Anything (Le et al., 2024) assembles articulated assets through VLM-generated programs and visual feedback; PAct (Liu et al., 2026) generates parts from images and masks and estimates joints from denoising features. Beyond part decomposition and articulation, ϕ-Art predicts material attributes and explicit spatial layouts to guide downstream synthesis. We retain frozen OmniPart for detailed meshes, placing our contribution in image-grounded semantic, spatial, and physical prediction rather than a new part-mesh backbone.

2.3 Physical 3D Generation

Physical 3D generation addresses both geometric compatibility and physical parameterization. PhysComp (Guo et al., 2024) optimizes geometry under prescribed materials and equilibrium constraints, while PhyCAGE (Yan et al., 2024) reduces interpenetration through simulation-guided optimization. DSO (Li et al., 2025a) aligns generators with stability feedback; LLM-to-Phy3D (Wong et al., 2025) optimizes prompts using visual and fluid-dynamics evaluations. PhysX-3D (Cao et al., 2025a) introduces PhysXNet and coupled geometry–physics latent generation, whereas PhysX-Anything (Cao et al., 2025b) autoregressively predicts physical descriptions and part-wise voxels before geometric refinement. Our distinction is not joint geometry–physics modeling alone, which these methods already address. Instead, ϕ-Art separates structured prediction from coarse geometry decoding, predicting whole-object codes in parallel rather than through iterative flow sampling or autoregressive geometry generation. Structured-context corruption further encourage visual grounding. Table 1 summarizes the task-level capabilities of representative methods.

3 Method

As shown in Fig. 2, ϕ-Art queries a VLM twice: first to predict part semantics, spatial layouts, articulation, and physical attributes, then to obtain a compressed geometry token (Section 3.1). A query-based decoder fuses its hidden state with image features to predict whole-object geometry codes in parallel (Section 3.2). Bounding-box corruption and silhouette supervision encourage visual grounding during training (Section 3.3). The decoded occupancy, predicted 3D boxes, input image, and visible-part masks condition a frozen OmniPart generator (Yang et al., 2025) to synthesize textured part meshes. These meshes are assembled using the predicted kinematic structure and associated with physical attributes to form the final articulated asset (Section 3.4).

3.1 Structured Object Representation

Given an image I, we predict a structured object description

$$S = \left(d, \{(s_i, b_i, v_i, a_i)\}_{i=1}^{P}, \mathcal{G}\right), \tag{1}$$

where d denotes object dimensions, and each of the P parts has a semantic label s_i, a 3D bounding box b_i, a visible-image bounding box v_i, and physical attributes a_i. These attributes include material, density, Young's modulus, Poisson's ratio, and affordance priority. The kinematic graph $\mathcal{G}$ specifies

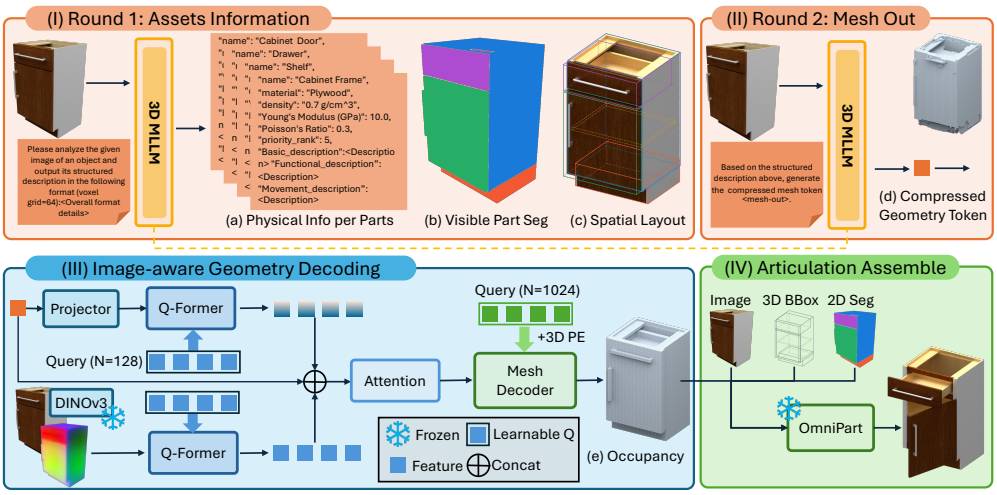

Figure 2: Overview of ϕ-Art. We query the 3D MLLM twice. (I) At Round 1, it predicts structured part information, including semantics, physical attributes, and spatial layouts. (II) A second round produces a compressed geometry token `<mesh-out>`. (III) Q-Formers aggregate this geometry token and image features for the geometry decoding. (IV) The occupancy, predicted 3D boxes, input image, and part masks condition a frozen OmniPart generator to synthesize the final output.

part grouping and hierarchy, motion types, joint axes or pivots, and motion limits. The 3D boxes provide a coarse part layout in a shared canonical frame, while d specifies absolute scale.

Following PhysX-Anything (Cao et al., 2025b), we serialize S as structured text. A VLM initialized from ShapeLLM-Omni (Ye et al., 2025) first predicts S autoregressively. A second query, conditioned on I and S, produces a compressed geometry token, `<mesh-out>`. Its hidden state h_g and image features jointly condition the parallel decoder to predict whole-object geometry codes.

3.2 Parallel Geometry Decoding

Multimodal conditioning. Learned conditioning queries cross-attend to a linear projection of h_g through a Q-Former-style transformer. A separate set of image queries aggregates patch features from DINOv3 encoder. A fusion layer updates the conditioning queries using their current representations, h_g, and the aggregated image features, combining structured object context with visual evidence.
Parallel code prediction. A transformer decodes N geometry queries from the fused conditioning tokens in parallel. Each query combines a learned embedding with positional embeddings of its three spatial indices, following the pretrained tokenizer's layout. Each output is projected to $u_j \in \mathbb{R}^C$ and matched against the frozen VQ-VAE codebook $E = \{e_k\}_{k=1}^K$:

$$p_j(k) = \mathrm{softmax}_k\left(\frac{\langle u_j, e_k\rangle}{\tau\|u_j\|_2\|e_k\|_2}\right), \qquad \hat{z}_j = \arg\max_k p_j(k), \tag{2}$$

where C is the embedding dimension and τ is a temperature. The frozen VQ-VAE decoder D maps the selected code embeddings to whole-object occupancy logits, whose positive entries define the predicted occupied support $\hat{V}$.
Geometry supervision. Given reference code indices $z_j^\star$, we supervise both code classification and embedding regression:

$$\mathcal{L}_{\text{code}} = -\frac{1}{N}\sum_{j=1}^{N}\log p_j(z_j^\star), \qquad \mathcal{L}_{\text{emb}} = \frac{1}{NC}\sum_{j=1}^{N}\|u_j - e_{z_j^\star}\|_2^2. \tag{3}$$

These losses update the geometry decoder and the VLM through h_g, while the codebook and occupancy decoder remain frozen.

3.3 Visual Grounding During Training

We combine structured-context corruption with silhouette supervision to reduce reliance on textual layouts and encourage geometry prediction from visual evidence.

Table 2: Quantitative comparison with existing methods on the full 1,310-object benchmark.

	Geometry			Physical & Articulation Attributes				
Method	F-score↑	CD↓	PSNR↑	Abs. Scale↓	Material↑	Affordance↑	Kinematics (VLM)↑	Description↑
URDFormer	14.66	347.99	7.57	–	–	–	0.31	–
URDF-Anything+	25.51	290.84	–	–	–	–	0.45	–
Articulate-Anything	32.35	236.38	12.20	–	–	–	0.43	–
PAct (SAM3 masks)	77.85	74.69	15.01	–	–	–	0.53	–
PAct	80.65	68.49	**15.24**	–	–	–	0.73	–
PhysX-Anything	72.67	82.08	15.10	9.91	**12.47**	**13.73**	**0.74**	**17.26**
ϕ-Art (SAM3 masks)	82.43	57.85	14.35	**8.62**	12.40	13.60	0.62	16.19
ϕ-Art	**82.44**	**57.84**	14.35	**8.62**	12.40	13.60	0.68	16.19

Structured-context corruption. Part boxes provide a coarse spatial plan, but can allow geometry prediction to bypass visual evidence. We corrupt the structured description S through two complementary operations. Specifically, *Bounding-box corruption* perturbs 3D box coordinates, drops individual box entries, and occasionally applies global scaling and translation to the layout. *Field masking* independently masks object names, categories, dimensions, part descriptions, visible-image boxes, and grouping information. Both operations preserve field headers and the surrounding serialization. The corrupted description $\tilde{S}$ conditions the second VLM query that produces the geometry-conditioning state h_g. The image and geometry target remain unchanged, encouraging geometry prediction from visual evidence when textual guidance is unreliable. We mix clean and corrupted contexts during training and exclude corrupted values and masked spans from $\mathcal{L}_{\text{LM}}$, retaining supervision on unmodified response tokens. Corruption is disabled at inference.

Silhouette supervision. We supervise decoded occupancy against the input foreground mask. A straight-through estimator preserves hard code selection while passing gradients through the predicted embeddings: $z_j^{\text{ST}} = \tilde{u}_j + \text{sg}\left(e_{\hat{z}_j} - \tilde{u}_j\right)$, where sg stops gradients and $\tilde{u}_j$ rescales u_j toward the selected code's norm with a bounded gain. We obtain occupancy probabilities $o = \sigma(D(z^{\text{ST}}))$ and project them along the known training camera rays:

$$\hat{m}_p = \max_{t \in \mathcal{T}_p} o(r_p(t)), \qquad \mathcal{L}_{\text{sil}} = \mathcal{L}_{\text{WBCE}}(\hat{m}, m^\star) + \mathcal{L}_{\text{Dice}}(\hat{m}, m^\star), \tag{4}$$

where $\mathcal{T}_p$ contains samples along ray r_p, $o(r_p(t))$ is trilinearly interpolated, and $m^\star$ is the reference foreground mask. The loss combines weighted binary cross-entropy and Dice loss. For this supervision, we recompute the geometry decoder with detached VLM states, so silhouette gradients update only the geometry decoder.

Training objective. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{code}} + \lambda_{\text{emb}}\mathcal{L}_{\text{emb}} + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}}, \tag{5}$$

where $\mathcal{L}_{\text{LM}}$ is cross-entropy over response tokens, and λ_{emb} and λ_{sil} weight the auxiliary losses.

3.4 Part Synthesis and Articulated Assembly

Part synthesis. A frozen OmniPart generator (Yang et al., 2025) takes the input image I, predicted occupancy $\hat{V}$, part boxes $\{b_i\}$, and visible-part masks M to jointly synthesize textured part meshes in a shared object frame. The boxes define part slots by selecting occupied coordinates within each region, while the masks provide decomposition cues without requiring one-to-one correspondence with the boxes. Our automatic pipeline obtains these masks by prompting SAM3 with the predicted image-space boxes $\{v_i\}$. The main evaluation uses ground-truth masks for controlled comparisons.

Articulated assembly. We assemble the generated parts according to the predicted kinematic graph $\mathcal{G}$. A deterministic exporter writes URDF and MJCF descriptions with the meshes and predicted joint parameters. For MJCF, the predicted dimensions d determine global scale, and part densities are assigned to the corresponding meshes. The remaining semantic and physical attributes are retained in the structured description S.

4 Experiments

4.1 Experimental Setup

Datasets and splits. We follow the PhysX-Anything data construction and train on the filtered PhysXNet and Mobility training sets. After object-level filtering, the training set contains 31,412

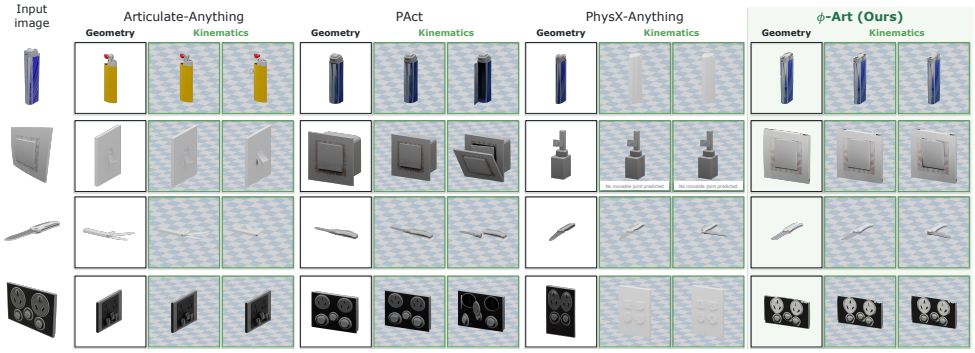

Figure 3: Qualitative Results of Asset Generation and Articulation Generation.

objects and 25 rendered views per object, yielding 785,300 image–object pairs. We remove objects containing more than 50 parts or an excessively long structured response, and use a unified 64-grid convention for coarse geometry, part boxes, and articulation parameters. The main benchmark contains the complete filtered test sets: 955 objects from PhysXNet and 355 objects from Mobility, for a total of 1,310 test objects. A fixed stratified subset of 435 objects, containing 317 PhysXNet and 118 Mobility objects, is used for scaling studies and controlled ablations.

Implementation details. We initialize the VLM from ShapeLLM-7B-omni (Ye et al., 2025) and fine-tune the language backbone and multi-modal projector with full-parameter updates. The VLM learning rate is 10^{-5}, while the geometry decoder uses 3×10^{-5}. Training uses four H200 GPUs with an effective batch size of 32, 500 warm-up steps, and cosine decay to 0.1 of the peak learning rate. We train the model for five epochs, with DINOv3 image encoder, and official OmniPart checkpoint frozen. OmniPart is kept identical across all experiments involving our method. All result-bearing evaluations use textured final meshes.

Baselines and inference inputs. We compare with PhysX-Anything (Cao et al., 2025b), PAct (Liu et al., 2026), Articulate-Anything (Le et al., 2024), URDF-Anything+ (Wu et al., 2026), and URDFormer (Chen et al., 2024). We reproduce PhysX-Anything using our training split before evaluation while for the rest methods we retain their released pretrained models and official inference pipelines. For comparison, we follow the evaluation setting and provide ground-truth 2D part masks to the downstream part-generation stage, consistent with OmniPart, PAct and Physx-Anything. We additionally evaluate a fully automatic variant of ϕ-Art, where the visible part masks required by OmniPart are obtained from our predicted 2D part boxes using SAM3.

4.2 Evaluation Metrics

Geometry and physical attributes. Following prior work, we evaluate geometry using F-score, Chamfer distance (CD), and novel-view RGB PSNR. Physical properties are evaluated with absolute-scale error and PSNR on rendered material, affordance, and description maps. We additionally report the VLM-based kinematic score from the PhysX-Anything protocol to measure articulated motion quality. Detailed metric definitions and failure handling are provided in the supplementary material.

Part-level spatial prediction. We evaluate Round 1 3D part layout using BBox IoU together with set-level $F@\tau$ and Hungarian-matched 1:1 $F@\tau$. For 2D grounding, predicted visible-part boxes are used to prompt SAM3, and the resulting masks are evaluated with mask IoU and mask F-score. Detailed metric definitions are provided in the supplementary material.

4.3 Main Results

Table 3: 3D part-layout prediction on the full 1,310-object test set.

Method	BBox IoU↑	$F@0.25$ ↑	$F@0.50$ ↑	1:1 $F@0.25$ ↑	1:1 $F@0.50$ ↑	1:1 $F@0.75$ ↑
TRELLIS + PartField	21.538	36.878	16.609	30.938	14.883	5.925
OmniPart BBoxGen	24.446	37.192	19.976	33.172	18.517	7.446
ϕ-Art	**32.688**	**51.940**	**31.836**	**50.003**	**31.302**	**12.362**

Comparison with prior methods. Table 2 reports the primary comparison on the complete 1,310-object benchmark. The table follows the PhysX-Anything metric family so that geometry, physical

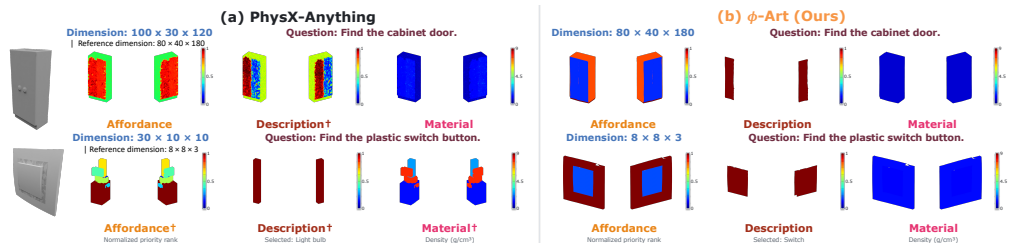

Figure 4: Qualitative results of physical attributes. Given a single real-world image as input,ϕ-Art produces high-quality physical attributes across diverse object categories

attributes, and articulation are evaluated under a common protocol. Detailed results on PhysXNet and Mobility are provided in Table 8 in the appendix.

As shown in Table 2, ϕ-Art achieves the strongest geometry reconstruction performance among the compared articulated and physics-aware simulation-ready asset generation methods, while remaining competitive in physical-attribute and kinematic prediction, Figure 3 and Figure 4. This indicates that our formulation effectively improves geometric fidelity while preserving the strong physical reasoning capability of the VLM.

Since the part-synthesis stage requires visible-part segmentation as conditioning, we additionally evaluate a fully automatic setting in which ground-truth masks are replaced by masks obtained from our Round 1 2D visible-part grounding and SAM3. We similarly evaluate PAct with automatically generated masks. As shown in Table 2, ϕ-Art is highly robust to this replacement: geometry and physical-attribute performance remain nearly unchanged, while most of the kinematic performance is retained. In contrast, PAct exhibits a more pronounced degradation in geometry and articulation. This demonstrates both the robustness of our part-synthesis pipeline to imperfect mask conditioning and the effectiveness of our 2D visible-part grounding for fully automatic asset generation.

3D part layout prediction. A key role of VLM stage is to recover the spatial layout of articulated parts before detailed part geometry is synthesized. We evaluate the predicted set of 3D part bounding boxes on the full 1,310-object test set. We compare our Round 1 prediction with the bounding-box generation pipeline used by OmniPart and a TRELLIS+PartField baseline. Since the predicted and ground-truth parts form unordered sets, we report both permissive set-level matching metrics and stricter one-to-one matching metrics.

As shown in Table 3, ϕ-Art consistently achieves the best performance across all 3D part-layout metrics. The improvement persists under strict one-to-one matching, demonstrating that our model captures not only the approximate spatial extent of object parts, but also their correct decomposition and relative arrangement. This stronger 3D structural understanding provides a more reliable conditioning signal for downstream part synthesis.

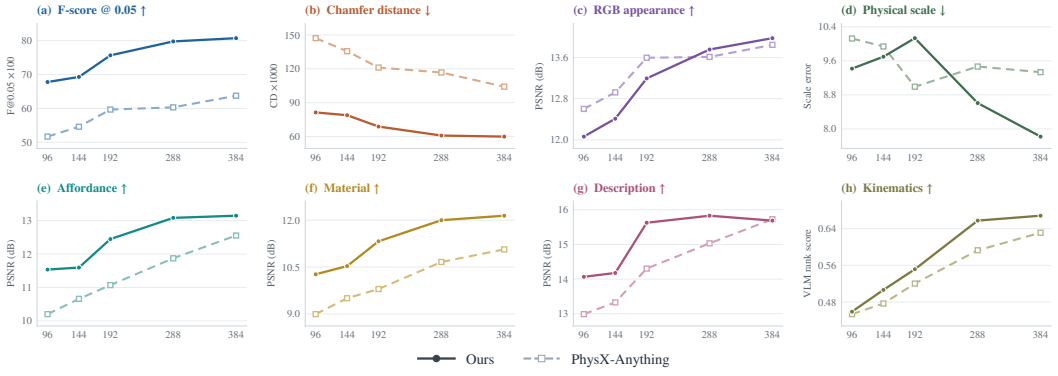

Figure 5: **Training-compute scaling against PhysX-Anything.** ϕ-Art consistently learns stronger geometric representations with increasing compute, while also exhibiting favorable scaling in appearance, physical attributes, and kinematics.

2D part grounding. We further evaluate whether the visible-part boxes predicted by Round 1 provide reliable image-space grounding for downstream part segmentation. We compare automatic

Table 6: Scaling with the number of training objects, evaluated on the fixed 435-object subset.

	Geometry			Physical & Articulation Attributes				
Training objects	F-score↑	CD↓	RGB PSNR↑	Abs. Scale↓	Material PSNR↑	Affordance PSNR↑	Kinematics (VLM)↑	Description PSNR↑
30%	63.624	94.114	11.793	8.776	10.034	11.426	0.474	14.124
60%	78.320	64.990	13.523	9.489	11.315	12.590	0.706	14.955
100%	**81.922**	**58.459**	**14.144**	**8.334**	**12.195**	**13.470**	**0.812**	**16.037**

SAM3 segmentation without spatial prompts, an oracle setting using ground-truth part boxes, and our fully automatic setting using visible-part boxes predicted by Round 1.

As shown in Table 4, our predicted boxes substantially improve part segmentation over prompt-free SAM3, recovering roughly 60% of the performance gap between direct segmentation and the oracle GT-box setting. This demonstrates that Round 1 provides accurate image-space localization of visible parts, enabling reliable mask generation for fully automatic downstream part synthesis.

Table 4: Evaluation of 2D grounding.

Setting	Mask IoU↑	Mask F-score↑
SAM3 direct	41.0	42.8
GT boxes + SAM3	**82.7**	**91.1**
Ours boxes + SAM3	65.4	71.3

4.4 Efficiency and Scaling

Inference efficiency. We compare inference latency with PhysX-Anything under the same hardware and output setting. As shown in Table 5, our Round 1 is nearly 20× faster, reflecting the efficiency of parallel geometry decoding over long autoregressive geometry-token generation. Despite the expensive downstream part-synthesis, our pipeline reaches an overall speedup of about 1.2×.

Table 5: Stage-wise and end-to-end inference latency of ϕ-Art and PhysX-Anything under matched hardware.

Method	Round 1(s)↓	Round 2(s)↓	End-to-End(s)↓
PhysX-Anything	135.97	**75.89**	211.86
ϕ-Art	**7.27**	169.03	**176.03**

Scalability of Compute. Intermediate training checkpoints are evaluated at increasing budgets up to 384 H200-hours, using the same evaluation set and protocol. As shown in Figure 5, ϕ-Art consistently exhibits more favorable scaling across geometry and physical-attribute metrics. In particular, geometric reconstruction improves rapidly with additional compute Similar improvements are observed for appearance, physical attributes, and kinematics. Under matched compute, our model consistently achieves stronger geometry reconstruction and competitive or better physical prediction across the evaluated budgets. The reproduced PhysX-Anything checkpoint used in Table 2, in comparison, requires roughly 2,800 H200-hours of training to reach its reported performance. These results indicate that the advantage of our formulation is not tied to a particular checkpoint: it learns substantially more efficiently and continues to benefit from additional training compute.

Scalability of Data. Finally, we examine whether ϕ-Art continues to benefit from a larger and more diverse training set. We train the same model using nested subsets containing 30%, 60%, and 100% of the available training objects, while keeping the other settings unchanged. As shown in Table 6, increasing the amount of training data consistently improves geometrics and physics. The gains remain visible from 60% to the full training set, suggesting that the model has not saturated at the current data scale and can continue to benefit from additional simulation-ready training assets.

4.5 Ablation Studies

We ablate 3 key components of ϕ-Art on the fixed 435-object subset: the parallel Q-Former geometry decoder, BBox corruption during structured prediction, and silhouette supervision for geometry decoding. The *w/o Q-Former* variant directly generates geometry tokens autoregressively from the VLM, while the other two variants remove the visual-grounding supervision from the full model.

As shown in Table 7, all three components contribute substantially to the final performance. Removing the Q-Former leads to a Removing the Q-Former causes a clear drop in geometry fidelity, while removing BBox corruption hurts geometry, kinematics, and physical understanding, highlighting the role of image-grounded spatial reasoning. Silhouette supervision further improves geometric fidelity and downstream physical-attribute prediction. The full model achieves the best performance across the reported metrics.

The qualitative comparison in Figure 6 further illustrates these effects. Without the Q-Former or BBox corruption, the generated assets exhibit larger structural discrepancies and less coherent articulated

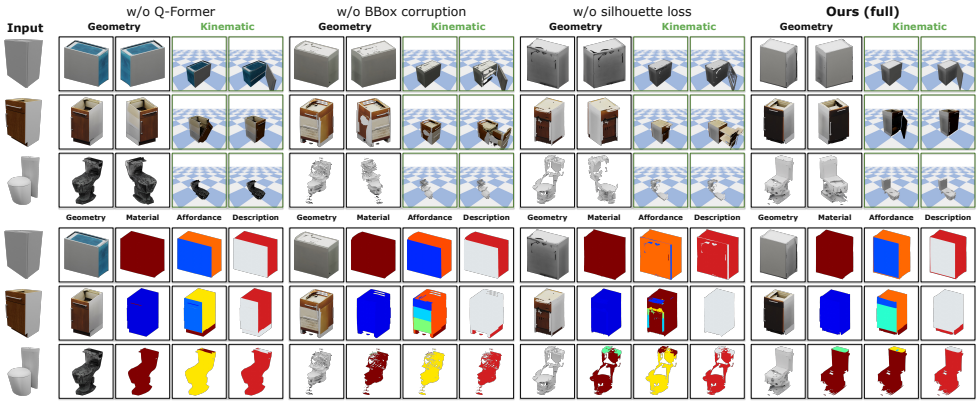

Figure 6: **Qualitative ablation of the main design components.** For each variant, we visualize reconstructed geometry, articulated motion, and the predicted material, affordance, and description.

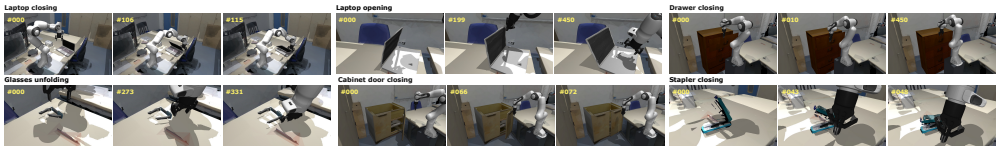

Figure 7: Robot Manipulation on Generated sim-ready 3D assets of ϕ-Art. The results show that our generated sim-ready assets exhibit highly physically plausible behavior and accurate geometric structure across diverse tasks. (Zoom in for better details.)

motion, while removing silhouette supervision introduces visible geometry artifacts. Combining all three components yields more faithful object geometry, cleaner articulation, and more consistent physical-property predictions.

Table 7: Ablation of geometry decoding and visual-grounding components on 435 objects.

	Geometry			Physical & Articulation Attributes				
Variant	F-score↑	CD↓	RGB PSNR↑	Abs. Scale↓	Material PSNR↑	Affordance PSNR↑	Kinematics (VLM)↑	Description PSNR↑
w/o Q-Former	64.360	79.721	12.580	9.284	10.138	11.225	0.651	13.104
w/o BBox corruption	68.232	84.063	12.234	9.031	10.391	11.710	0.471	14.578
w/o silhouette loss	73.346	68.706	12.912	9.015	11.289	12.445	0.623	14.779
Ours	**81.922**	**58.459**	**14.144**	**8.334**	**12.195**	**13.470**	**0.736**	**16.037**

Simulation-ready Evidence. We use ϕ-RIE (Yang et al., 2026) to integrate our generated articulated assets into reconstructed environments for robot interaction. Figure 7 shows laptop opening and closing, cabinet door and drawer closing, glasses unfolding, and stapler closing. These examples involve both rotational and translational part motion, providing qualitative evidence that the generated geometry and kinematic structure support articulated manipulation within a scene.

5 Conclusion

We presented ϕ-Art, a framework for generating physics-aware articulated assets from a single image. A VLM provides a coarse structural and physical plan, while a Q-Former-style decoder combines its hidden representations with image features to predict whole-object geometry codes in parallel. Bounding-box corruption and silhouette supervision encourage the decoder to use visual evidence rather than rely on textual layouts alone. The resulting shape prior and part layouts guide textured part synthesis and kinematic assembly. Experiments show improved geometric fidelity and part-layout prediction, with competitive physical-attribute and kinematic performance. Compared with PhysX-Anything, ϕ-Art also achieves greater training efficiency and faster end-to-end inference. Robot interaction examples further demonstrate articulated manipulation with the generated assets in reconstructed environments. These results support combining explicit structural planning with image-grounded geometry generation to produce articulated assets for interactive simulation.

AI USE STATEMENT

We used generative AI tools to assist model development and implementation, including code generation, and to improve the clarity and readability of the manuscript. Our use of these tools was limited to these purposes. All AI-generated code was reviewed and checked for correctness by at least two authors. We reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims, code, data, figures, and other artifacts.

REFERENCES

Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. 2

Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-3d: Physical-grounded 3d asset generation. *arXiv preprint arXiv:2507.12465*, 2025a. 1, 3

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. 1, 3, 4, 6

Zoey Chen, Aaron Walsman, Marius Memmel, Kaichun Mo, Alex Fang, Karthikeya Vemuri, Alan Wu, Dieter Fox, and Abhishek Gupta. Urdformer: A pipeline for constructing articulated simulation environments from real-world images. *arXiv preprint arXiv:2405.11656*, 2024. 6

Minghao Guo, Bohan Wang, Pingchuan Ma, Tianyuan Zhang, Crystal Elaine Owens, Chuang Gan, Joshua B. Tenenbaum, Kaiming He, and Wojciech Matusik. Physically compatible 3d object modeling from a single image. *arXiv preprint arXiv:2405.20510*, 2024. 3

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. 13

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. 1

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. 1, 3, 6

Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Dso: Aligning 3d generators with simulation feedback for physical soundness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6772–6783, 2025a. 3

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b. 2

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

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. 3, 6

David Marr and Herbert Keith Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, pp. 269–294, 1978. 2

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 909–918, June 2019. 3

David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3):353–383, 1977. 2

Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. MeshGPT: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19615–19625, 2024. 2

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2

Jiaxiang Tang, Ruijie Lu, Zhaoshuo Li, Zekun Hao, Xuan Li, Fangyin Wei, Shuran Song, Gang Zeng, Ming-Yu Liu, and Tsung-Yi Lin. Efficient part-level 3D object generation via dual volume packing. In *Advances in Neural Information Processing Systems*, 2025. 3

Melvin Wong, Yueming Lyu, Thiago Rios, Stefan Menzel, and Yew-Soon Ong. Llm-to-phy3d: Physically conform online 3d object generation with llms. *arXiv preprint arXiv:2506.11148*, 2025. 3

Zhuangzhe Wu, Yue Xin, Chengkai Hou, Minghao Chen, Yaoxu Lyu, Jieyu Zhang, and Shanghang Zhang. Urdf-anything+: Autoregressive articulated 3d models generation for physical simulation. *arXiv preprint arXiv:2603.14010*, 2026. 6

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11097–11107, 2020. 3

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. 1, 2

Han Yan, Mingrui Zhang, Yang Li, Chao Ma, and Pan Ji. Phycage: Physically plausible compositional 3d asset generation from a single image, 2024. 3

Runyi Yang, Deheng Zhang, Xiaoye Wang, Kanzhi Wu, Lei Sun, Ajad Chhatkuli, Kunyu Peng, Luc Van Gool, and Danda Pani Paudel. ϕ-RIE: From photorealistic reconstruction to interactive environments, 2026. 9

Yunhan Yang, Yufan Zhou, Yuan-Chen Guo, Zi-Xin Zou, Yukun Huang, Ying-Tian Liu, Hao Xu, Ding Liang, Yan-Pei Cao, and Xihui Liu. Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025. 1, 2, 3, 5

Junliang Ye, Zhengyi Wang, Ruowen Zhao, Shenghao Xie, and Jun Zhu. Shapellm-omni: A native multimodal llm for 3d generation and understanding. *arXiv preprint arXiv:2506.01853*, 2025. 2, 4, 6, 13

Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3DShape2VecSet: A 3D shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592442. 2

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. CLAY: A controllable large-scale generative model for creating high-quality 3D assets. *ACM Transactions on Graphics*, 43(4):1–20, 2024. doi: 10.1145/3658146. 2

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 *Advances in Neural Information Processing Systems*, 2023. 2

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, Huiwen Shi, Sicong Liu, Junta Wu, Yihang Lian, Fan Yang, Ruining Tang, Zebin He, Xinzhou Wang, Jian Liu, Xuhui Zuo, Zhuo Chen, Biwen Lei, Haohan Weng, Jing Xu, Yiling Zhu, Xinhai Liu, Lixin Xu, Changrong Hu, Shaoxiong Yang, Song Zhang, Yang Liu, Tianyu Huang, Lifu Wang, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Chao Zhang, Yonghao Tan, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu, Zhichao Hu, Lei Qin, Jianbing Peng, Zhan Li, Minghui Chen, Xipeng Zhang, Lin Niu, Paige Wang, Yingkai Wang, Haozhao Kuang, Zhongyi Fan, Xu Zheng, Weihao Zhuang, YingPing He, Tian Liu, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, Jingwei Huang, and Chunchao Guo. Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation, 2025. arXiv v5 (2026-05-14). 2

A Additional Details

The following appendices provide additional technical details and experimental results that support the main findings of this work.

A.1 More Implementation Details

Geometry decoder. The geometry decoder predicts $N = 1{,}024$ codes using a frozen VQ-VAE codebook with $K = 8{,}192$ entries of dimension $C = 32$ (Ye et al., 2025). We use 128 conditioning queries and 64 image queries. The conditioning transformer and geometry transformer each contain four layers, while the image aggregation and multimodal fusion transformers each contain one layer. These transformer blocks use a hidden dimension of 1,024, 16 attention heads, a feed-forward dimension of 4,096, and GELU activations (Hendrycks & Gimpel, 2023). Geometry queries use learned positional embeddings along the three axes of a $16 \times 16 \times 4$ code layout. Cosine code classification uses a temperature of $\tau = 0.07$. The frozen VQ-VAE decoder produces occupancy logits on a 64^3 grid.

Visual features and training configuration. We extract image features from a frozen DINOv3 ViT-L/16 encoder at an input resolution of 224×224. Patch features from transformer blocks with zero-based indices 4, 11, 17, and 23 are concatenated along the token dimension, retaining all 14×14 patches from each block. Training uses BF16 precision, a maximum sequence length of 5,120 tokens, a per-GPU batch size of two, and four gradient accumulation steps. The VLM visual encoder, DINOv3 encoder, VQ-VAE codebook, and occupancy decoder remain frozen.

Structured-context corruption. We quantize 3D bounding-box coordinates into 64 bins and visible-image bounding-box coordinates into 128 bins. Structured-context corruption is applied to 75% of training examples. For ordinary box perturbations, each 3D coordinate receives an independently sampled offset from $\{-4, -2, 0, 2, 4\}$ bins. Individual box entries are dropped with probability 0.1, while retaining at least one box. For 5% of corrupted examples, local jitter is replaced by a shared layout scaling factor sampled uniformly from $[0.6, 1.5]$ and per-axis translations sampled from $\{-12, -10, -8, 8, 10, 12\}$ bins. Coordinates are clipped to the valid range and box bounds are reordered when necessary. Within corrupted examples, each eligible field is independently masked with probability 0.3. The entire corrupted 3D bounding-box block and masked field contents are excluded from language-model supervision.

Silhouette supervision and inference. We set $\lambda_{\text{emb}} = \lambda_{\text{sil}} = 0.1$, with unit weights for language modeling and code classification. Silhouette supervision uses a 64×64 projection with 160 uniformly spaced samples per camera ray. Reference alpha masks are downsampled by area averaging, preserving soft foreground values. Camera rays are transformed into the canonical occupancy frame using precomputed normalization transforms. At inference, geometry codes are selected by argmax, and positive occupancy logits define occupied voxels. OmniPart synthesis uses 25 sampling steps and a guidance scale of 7.5.

A.2 Details of Evaluation Metrics

Geometry and appearance. Following the common protocol, we report novel-view RGB PSNR, Chamfer distance (CD), and shape F-score. For CD and F-score, predicted and reference rest-pose meshes are independently centered and normalized by their longest bounding-box side. We sample 100k surface points from each mesh and report symmetric Chamfer distance and F-score at the fixed distance threshold used throughout the benchmark. RGB PSNR is averaged over the same set of rendered novel views for all methods.

Physical attributes. We follow the PhysX-Anything evaluation protocol for physical attributes. Absolute scale measures object-scale error. Material, affordance, and description are evaluated as rendered property maps and reported using PSNR. We additionally report the VLM-based kinematic score used by the benchmark, which evaluates the similarity of the generated articulated motion to the reference motion.

Table 8: Detailed quantitative comparison on PhysXNet and Mobility.

	Geometry			Physical & Articulation Attributes				
Method	F-score↑	CD↓	PSNR↑	Abs. Scale↓	Material↑	Affordance↑	Kinematics (VLM)↑	Description↑
PhysXNet: 955 objects								
URDFormer	13.48	350.70	7.53	–	–	–	0.30	–
URDF-Anything+	24.44	294.21	–	–	–	–	0.44	–
Articulate-Anything	30.19	239.50	11.90	–	–	–	0.36	–
PAct (SAM3 masks)	79.60	70.04	14.95	–	–	–	0.51	–
PAct	83.12	62.32	**15.23**	–	–	–	**0.77**	–
PhysX-Anything	72.87	80.67	15.13	10.92	12.31	**13.78**	0.75	**16.76**
ϕ-Art (SAM3 masks)	85.23	52.58	14.63	**9.57**	12.39	13.73	0.65	15.71
ϕ-Art	**85.24**	**52.58**	14.63	**9.57**	**12.39**	13.73	0.71	15.71
Mobility: 355 objects								
URDFormer	17.69	341.06	7.69	–	–	–	0.34	–
URDF-Anything+	28.38	281.75	–	–	–	–	0.47	–
Articulate-Anything	38.20	227.95	13.01	–	–	–	0.63	–
PAct (SAM3 masks)	73.18	87.15	15.16	–	–	–	0.60	–
PAct	74.02	85.13	**15.25**	–	–	–	0.61	–
PhysX-Anything	72.13	85.88	15.03	7.19	**12.89**	**13.59**	**0.71**	**18.59**
ϕ-Art (SAM3 masks)	74.90	72.03	13.60	**6.06**	12.42	13.25	0.53	17.50
ϕ-Art	**74.92**	**71.99**	13.60	**6.06**	12.42	13.25	0.60	17.48

Missing predictions. To avoid favoring methods that fail to produce outputs on difficult examples, we compute submission-level metrics over the complete evaluation set rather than over valid predictions only. For a failed geometry prediction, its F-score is set to zero. For CD, we construct a failure surrogate by uniformly sampling points inside the axis-aligned bounding box of the ground-truth object and computing CD to the ground-truth surface point set under the same normalization; the resulting value is included in the average as usual. For missing rendered outputs, RGB, material, affordance, and description PSNR are set to 0 dB. All reported means therefore include failed objects with these predefined penalties.

3D part layout. We evaluate the 3D part layout predicted by round 1 on the full test set. Since predicted and ground-truth parts form unordered sets, we first compute the pairwise 3D box IoU matrix. For *BBox IoU*, Hungarian matching is used to maximize the total IoU between predicted and ground-truth boxes; the matched IoUs are summed and normalized by the number of ground-truth parts, such that unmatched ground-truth parts contribute zero.

We additionally report two complementary F-score protocols. For $F@\tau$, precision and recall independently test whether each predicted and ground-truth box has any counterpart with IoU $\geq \tau$, allowing multiple predictions to match the same part. In contrast, 1:1 $F@\tau$ first fixes the Hungarian one-to-one assignment and counts only matched pairs whose IoU exceeds τ, thereby more strongly penalizing duplicated and unmatched predictions. We report $F@0.25$, $F@0.50$, and 1:1 F at multiple IoU thresholds. All metrics are computed per object and macro-averaged over the complete test set, with failed cases assigned zero.

2D part grounding. We further evaluate whether the predicted visible-part boxes provide reliable image-space grounding for automatic part segmentation. Predicted 2D boxes are used to prompt SAM3, and the resulting masks are compared with ground-truth part masks using mask IoU and mask F-score. This evaluation isolates the quality of image grounding used by the fully automatic variant of our pipeline.

A.3 More Results

Dataset-wise breakdown of Table 2. Table 8 provides the PhysXNet and Mobility breakdown of the overall comparison reported in the main paper. The same overall trend holds across both subsets: ϕ-Art consistently provides stronger geometric reconstruction, while maintaining competitive physical-attribute and kinematic prediction. In particular, the geometry advantage is preserved on both PhysXNet and Mobility, indicating that the improvement is not driven by a single data source. The automatically grounded SAM3 variant also remains close to the ground-truth-mask setting on both subsets, further confirming the robustness of our downstream part synthesis to predicted part masks.

Output capabilities of compared methods. The compared methods differ in the scope of their generated asset representations. URDFormer, URDF-Anything+, Articulate-Anything, and PAct primarily focus on articulated geometry and kinematic structure, and do not predict the complete set of physical attributes evaluated in our benchmark. Accordingly, metrics such as absolute scale, material, affordance, and physical description are not applicable to these methods. PhysX-Anything and ϕ-Art instead generate a more complete simulation-ready representation that includes geometry, articulation, and physical properties.

Qualitative part grounding and spatial layout. Figure 8 provides additional qualitative results for the spatial predictions produced by Stage 1. Across diverse object categories, the SAM3 masks prompted by our predicted visible-part boxes closely follow the ground-truth part segmentation, while the predicted 3D bounding boxes capture the part decomposition and spatial arrangement of the reference objects. These examples complement the quantitative results in Tables 3 and 4, illustrating that our structured prediction stage provides consistent grounding in both image and 3D object space.

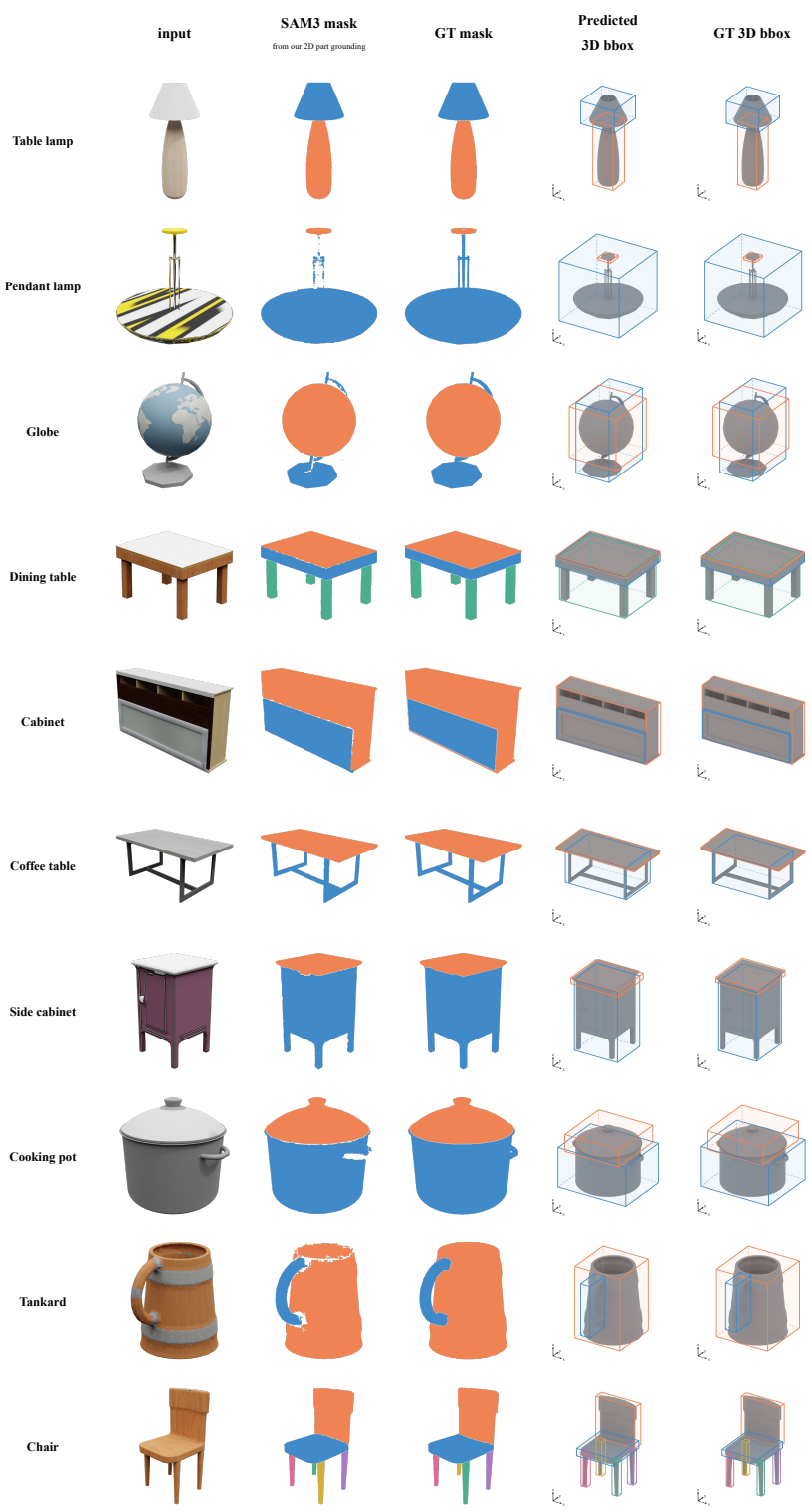

Figure 8: **Qualitative visualization of 2D part grounding and 3D part-layout prediction.** From top to bottom, we show the input image, SAM3 masks prompted by our predicted 2D visible-part boxes, ground-truth visible-part masks, predicted 3D part bounding boxes, and ground-truth 3D part bounding boxes.