PhysX-Omni: Unified Simulation-Ready Physical 3D Generation for Rigid, Deformable, and Articulated Objects

Ziang Cao[1], **Yinghao Liu**[2], **Haitian Li**[1], **Runmao Yao**[1], **Fangzhou Hong**[1], **Zhaoxi Chen**[1], **Liang Pan**[2], **Ziwei Liu**[1]

[1] S-Lab, Nanyang Technological University, [2] ACE Robotics

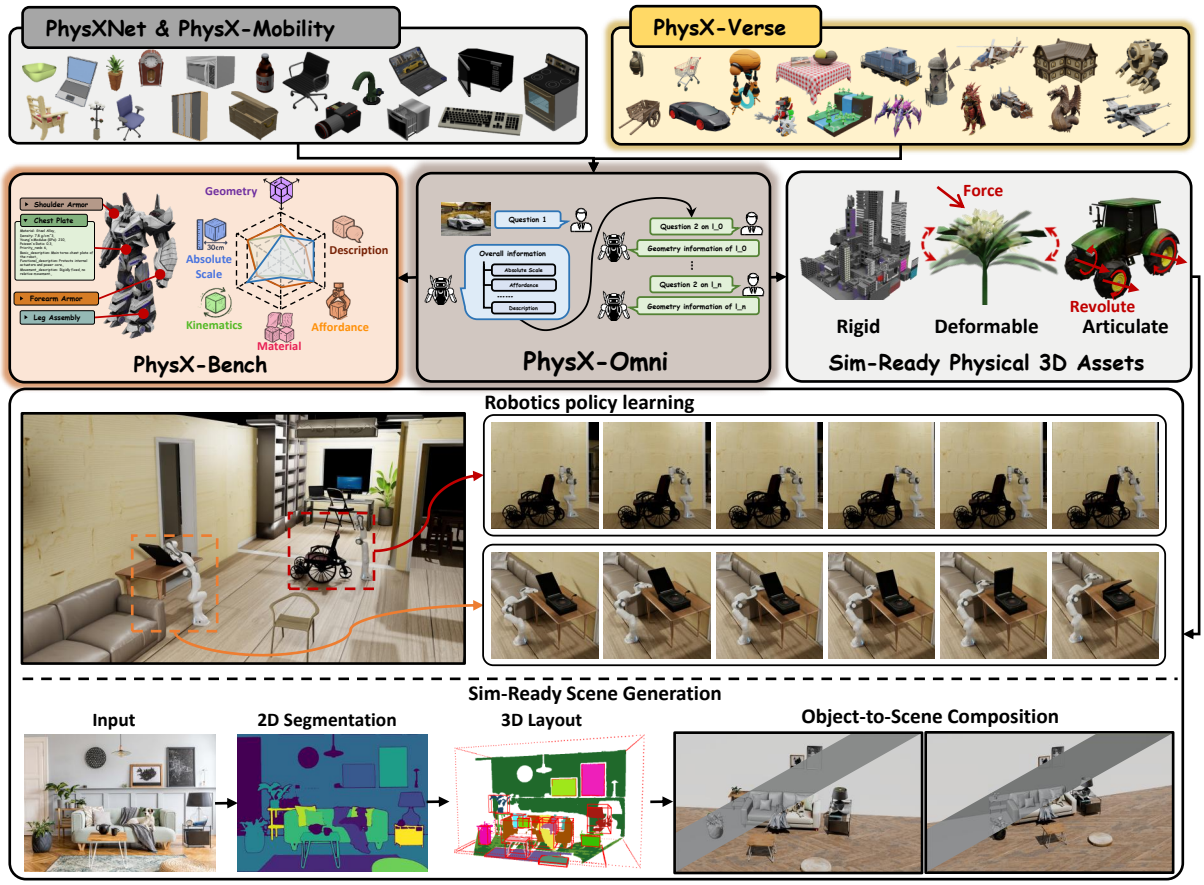

Figure 1: By exploiting the high diversity of PhysXVerse, PhysX-Omni is capable of generating detailed and general 3D assets covering rigid, deformable, and articulated objects, producing simulation-ready physical assets suitable for downstream applications.

Abstract

Simulation-ready physical 3D assets have emerged as a promising direction owing to their broad applicability in downstream tasks. However, most existing 3D generation methods either neglect physical properties or are limited to a single asset category, *e.g.*, rigid, deformable, or articulated objects. To address these limitations, we introduce **PhysX-Omni**, **a unified framework** for simulation-ready physical 3D generation across diverse asset types. Specifically, we develop a novel and efficient geometry representation tailored for Vision-Language Models, which di-

rectly encodes high-resolution 3D structures without compression, significantly improving generation performance. In addition, we construct the **first general simulation-ready 3D dataset**, **PhysXVerse**, covering diverse indoor and outdoor categories. Furthermore, to comprehensively and flexibly evaluate both generative and understanding capabilities in the wild, we propose **PhysX-Bench**, which encompasses six key attributes: geometry, **absolute scale**, **material**, **affordance**, **kinematics**, and **description**. Extensive experiments with conventional metrics and PhysX-Bench show that PhysX-Omni performs strongly in both generation and understanding. Moreover, additional studies further validate the potential of PhysX-Omni for applications in simulation-ready scene generation and robotic policy learning. We believe PhysX-Omni can significantly advance a wide range of downstream applications, particularly in embodied AI and physics-based simulation.

Official Page: https://physx-omni.github.io/
Correspondence: Ziwei Liu (ziwei.liu@ntu.edu.sg)

Contents

1 Introduction

High-quality simulation-ready (sim-ready) 3D assets have attracted significant attention due to their wide range of downstream applications in gaming design, robotics, embodied AI, and interactive simulation. However, most existing 3D generation approaches primarily focus on achieving photorealistic appearance and detailed geometric structures [1–8]. Despite their strong generative performance, the generated 3D assets often lack essential physical attributes required for real-world deployment, thereby limiting their applicability, particularly in physics-based scenarios.

To bridge this gap, a number of works have focused on generating articulated assets [9–13] and deformable assets [14–19]. However, these methods typically model only a limited subset of physical attributes for a specific asset type (*e.g.*, articulated or deformable objects), while overlooking other essential properties. As pioneering efforts in sim-ready physical 3D generation [20, 21], they enable the synthesis of richer physical attributes. Nevertheless, they remain constrained by the scarcity of large-scale, high-quality annotated 3D datasets, which limits the diversity of generated assets and, consequently, their practical utility for downstream embodied AI and control tasks. Furthermore, the absence of effective benchmarks for evaluating physical attributes in real-world scenarios (without ground-truth annotations) significantly limits meaningful evaluation.

To address these challenges, we propose **PhysX-Omni**, a unified simulation-ready physical 3D generative framework that supports diverse object types, including rigid, deformable, and articulated assets , with broad potential applications as illustrated in Fig. 1. Specifically, we introduce a novel geometry representation tailored for Vision-Language Models (VLM), which directly models high-resolution 3D structures without requiring additional special tokens during training. By explicitly modeling 3D structure, PhysX-Omni avoid the failure modes caused by segmentation, thereby significantly improving generative performance. Moreover, since we avoid additional decoder refinement, our framework remains compatible with existing voxel-based decoders [1, 6, 22], enabling the synthesis of high-fidelity appearance.

To address data scarcity, we construct the first general simulation-ready physical 3D dataset, **PhysX-Verse**, which contains over 8K assets spanning more than 2K indoor and outdoor categories, *e.g.*, helicopters, tanks, racing cars, skyscrapers, and toys, curated and filtered from PartVerse [23]. Furthermore, to comprehensively evaluate simulation-ready 3D generation, we build the first physical 3D generative benchmark, **PhysX-Bench**, covering six key attributes: geometry, **absolute scale**, **material**, **affordance**, **kinematics**, and **description**. By leveraging physics-based simulation and powerful VLMs, PhysX-Bench enables robust and realistic evaluation in in-the-wild scenarios. Comprehensive experiments with conventional metrics and PhysX-Bench demonstrate that PhysX-Omni achieves superior performance in both generation quality and generalization compared to recent state-of-the-art methods. Finally, to validate deployability in standard simulators and physics engines, we conduct experiments in a common simulation environment, showing that our simulation-ready assets can be directly applied to contact-rich robotic policy learning. We believe our work opens up new opportunities for future research in 3D generation, embodied AI, and robotics.

To summarize, our main contributions are:

- We introduce **PhysX-Omni**, a novel unified framework for simulation-ready physical 3D generation across diverse asset types. By employing the new tailored geometry representation, our approach directly models detailed geometric structures, leading to significantly improvements in both performance and generalization.
- We construct the first general simulation-ready physical 3D dataset, **PhysXVerse**, covering over 2K indoor and outdoor categories (*e.g.*, trucks, jets, and flowers), with high-quality physical attribute annotations.
- We introduce the first benchmark for simulation-ready physical 3D generation, **PhysX-Bench**. By integrating physics-based simulation with powerful VLMs, PhysX-Bench provides a comprehen-

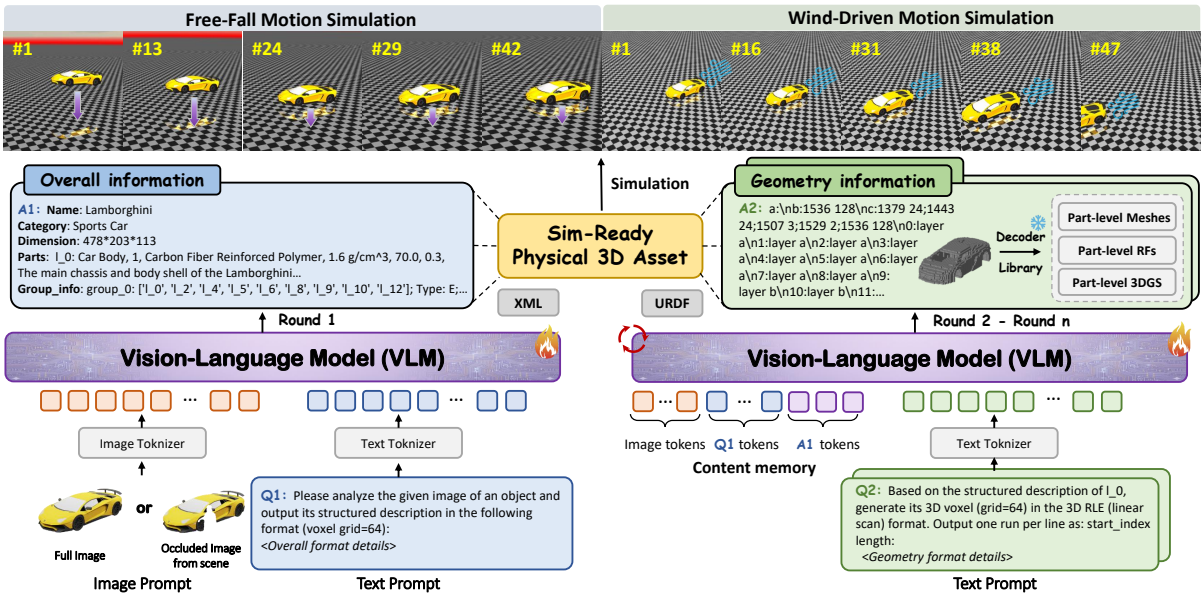

Figure 2: Given a single complete or partially occluded image, PhysX-Omni first infers high-level overall information. It then employs a multi-turn generation process to produce detailed part-level geometry. Owing to the inherent alignment between global and local representations, these outputs can be directly integrated into simulation-ready physical 3D assets.

sive and robust evaluation framework for assessing generation methods in real-world scenarios across six key attributes.

- Extensive evaluations on PhysX-Bench and conventional benchmarks demonstrate that PhysX-Omni achieves impressive generative quality and robust generalization. Moreover, we verify the deployability of our simulation-ready assets in standard simulation environments, facilitating downstream applications in embodied AI and robotic manipulation.

2 Related Works

2.1 Appearance-Centric 3D Generation

Early efforts in 3D generation were largely dominated by generative adversarial networks (GANs), which laid the foundation for this field [24, 25]. Despite their initial success, GAN-based approaches often suffer from instability and limited robustness when scaling to more complex and diverse data distributions. The introduction of DreamFusion [26] marked a significant shift by proposing score distillation sampling (SDS), which leverages the strong priors of pretrained 2D diffusion models. Nevertheless, such optimization-based pipelines remain computationally expensive and are prone to artifacts such as the Janus effect. To address these limitations, recent works increasingly favor feed-forward architectures, which offer improved efficiency and more stable generation behavior [1–3, 27–35]. In parallel, alternative paradigms have also been explored, including autoregressive approaches that model 3D structures sequentially [36, 37]. To mitigate the challenge of long token sequences in geometry modeling, LLaMA-Mesh [38] adopts a simplified mesh representation, while MeshLLM [4] introduces a hierarchical part-level generation strategy to further improve quality. ShapeLLM-Omni [5] instead compresses 3D representations via a VQ-VAE, but at the cost of introducing specialized tokens and a dedicated tokenizer, which complicates the training pipeline.

In contrast, PhysX-Anything [21] explores modeling simulation-ready physical 3D assets using pure text representations. Benefiting from the strong prior knowledge of VLMs, it achieves impressive generative performance and robustness. However, its reliance on an explicit segmentation stage introduces

a performance bottleneck, as the overall quality is constrained by the segmentation module. To overcome this limitation, we propose a new geometry representation that directly models high-resolution 3D structures. By simplifying the overall framework, our approach significantly improves generation performance over the baseline.

2.2 Physical 3D Asset Generation

Articulated object generation has recently gained increasing attention due to its broad range of downstream applications [13, 39–44]. Existing articulate generation approaches can be broadly categorized into several paradigms. A dominant line of work follows a retrieval-based strategy, where articulated assets are constructed by retrieving and assembling meshes from a predefined source library [9, 11]. While effective within known categories, such methods are inherently limited by the coverage of the database and struggle to generalize to novel structures. Another line of research adopts graph-structured representations [10, 45], integrating kinematic graphs with diffusion models to enable structure-aware generation. However, these approaches typically focus on geometry and lack the ability to produce high-quality textured assets, limiting their realism. Beyond these paradigms, optimization-based methods such as DreamArt [12] attempt to reconstruct articulated objects from video generation outputs. Despite their flexibility, they rely on manually annotated part masks and tend to become unstable when handling objects with many movable components. URDF-Anything [46] and URDF-Anything+ [47] directly generates URDF representations, but its performance heavily depends on high-quality point cloud inputs or mesh and it remains challenging to produce detailed textures. Recently, MonoArt [13] leverages priors from 3D generation and segmentation to infer kinematic parameters and achieve promising performance. Nevertheless, all those method primarily focuses on a single type of physical attribute and lacks a holistic modeling of physical objects. Beyond articulated object generation, several works have also explored modeling the deformation of 3D assets [16–19, 48]. However, these approaches also overlook other critical physical attributes, limiting their realism. To advance 3D generation toward physical fidelity, PhysXGen [20] introduces a unified framework that directly generates 3D assets with essential physical properties, such as absolute scale and density. Building upon this line of work, PhysX-Anything [21] further extends the paradigm to simulation-ready 3D asset generation. Nevertheless, it remains constrained by the limited diversity of available simulation-ready datasets and faces challenges in modeling high-quality, detailed assets efficiently.

To address these limitations, we propose a tailored geometry representation within a unified framework, along with the first general high-quality simulation-ready 3D dataset. Benefiting from both the enriched data diversity and the efficient geometry representation, our PhysX-Omni demonstrates strong robustness and superior performance in generating complex topologies and accurate physical attributes. We believe our approach opens up a promising direction for leveraging synthetic data to advance downstream applications.

3 Methodology

In this section, we describe the core components of PhysX-Omni, including the overall paradigm illustrated in Fig. 2, the newly constructed dataset, PhysXVerse, and the first benchmark for simulation-ready 3D assets, PhysX-Bench.

3.1 Generative paradigm of PhysX-Omni

PhysX-Omni adopts a VLM-based generation paradigm to produce simulation-ready physical assets through a coarse-to-fine global-to-local reasoning process, following [21]. As illustrated in Fig. 2, given a complete or partially occluded image, PhysX-Omni first performs holistic understanding to infer high-level global information, including the object category, semantic identity, absolute scale, component hierarchy, and potential physical properties. Such global understanding provides strong structural and

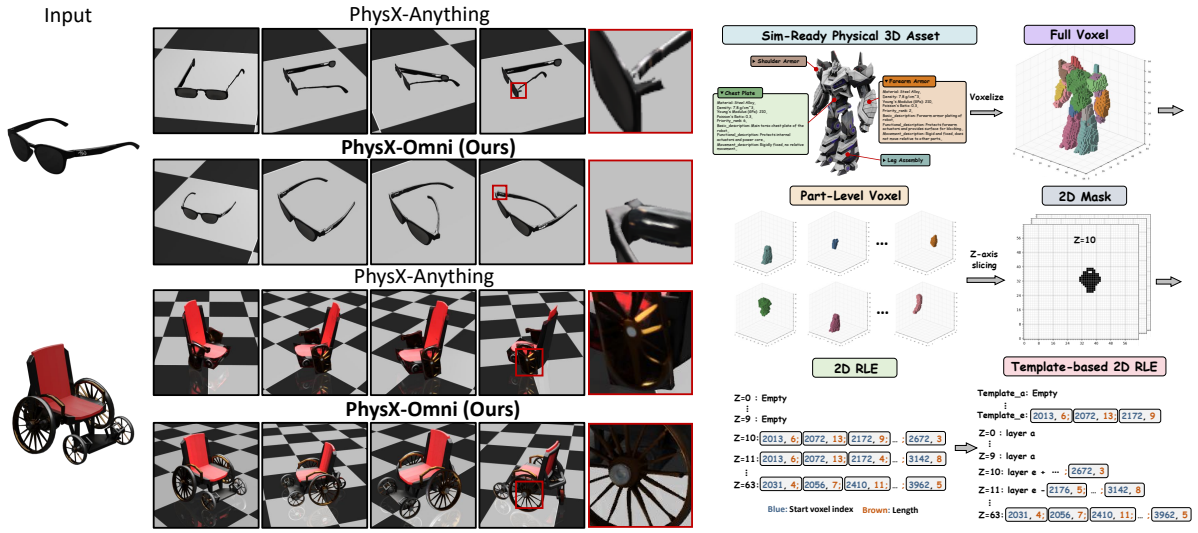

Figure 3: (a). Comparison of different geometry representations for 3D modeling. Leveraging the proposed geometry representation, PhysX-Omni effectively captures fine-grained 3D structures and enhances kinematic accuracy. (b). **Detailed geometry representation of our PhysX-Omni.** To directly model high-resolution 3D structures, we first slice part-level voxel grids along the z-axis. For each resulting 2D mask, we apply classical run-length encoding (RLE) to convert the binary image into a compact textual representation. To further improve compression efficiency, we introduce template layers, enabling other layers to be expressed as variations relative to templates.

semantic priors for subsequent part-level generation and helps maintain consistency between the overall object and its local components.

Based on the inferred global representation, PhysX-Omni further predicts the detailed geometric structure and physical attributes of each individual part. For the global representation, we follow the tree-structured and VLM-friendly formulation introduced in [20], which effectively organizes object-level and part-level information into a hierarchical representation compatible with autoregressive vision–language modeling.

For geometry representation, we introduce a novel high-resolution structure modeling strategy that directly encodes detailed 3D geometry in a compact and generation-friendly manner shown in Fig. 3b. Unlike prior methods that heavily rely on mesh decomposition or additional segmentation modules, our representation allows PhysX-Omni to directly model complex geometric structures while preserving explicit structural information. As a result, PhysX-Omni can seamlessly leverage a pre-trained voxel-based 3D decoder to generate high-quality meshes without requiring additional mesh segmentation processes, thereby significantly improving generation quality, robustness, and generalization ability, especially for objects with complex topologies and fine-grained structures.

Prior works have explored various compact 3D representations for vision–language modeling, including vertex quantization [4, 38], 3D VQ-GAN representations [5], and text-based voxel indices [21] to reduce sequence length and improve generation efficiency. However, these methods either rely on additional special tokens, suffer from limited geometric fidelity, or struggle to explicitly model high-resolution structures in a generation-friendly manner. To address these limitations while maintaining compatibility with existing VLM token spaces, we introduce a novel text-based geometry representation that does not require introducing additional special tokens into the language model vocabulary.

Specifically, inspired by classical 2D run-length encoding (RLE), we propose a template-based RLE representation to explicitly and directly model high-resolution 3D geometry. We first voxelize the simulation-ready assets and decompose them into part-level voxels according to the annotated object structure. Each part-level voxel is then sliced along the z-axis into a sequence of 2D binary masks. For each slice, we apply a compact 2D RLE formulation to encode the occupied regions into text tokens

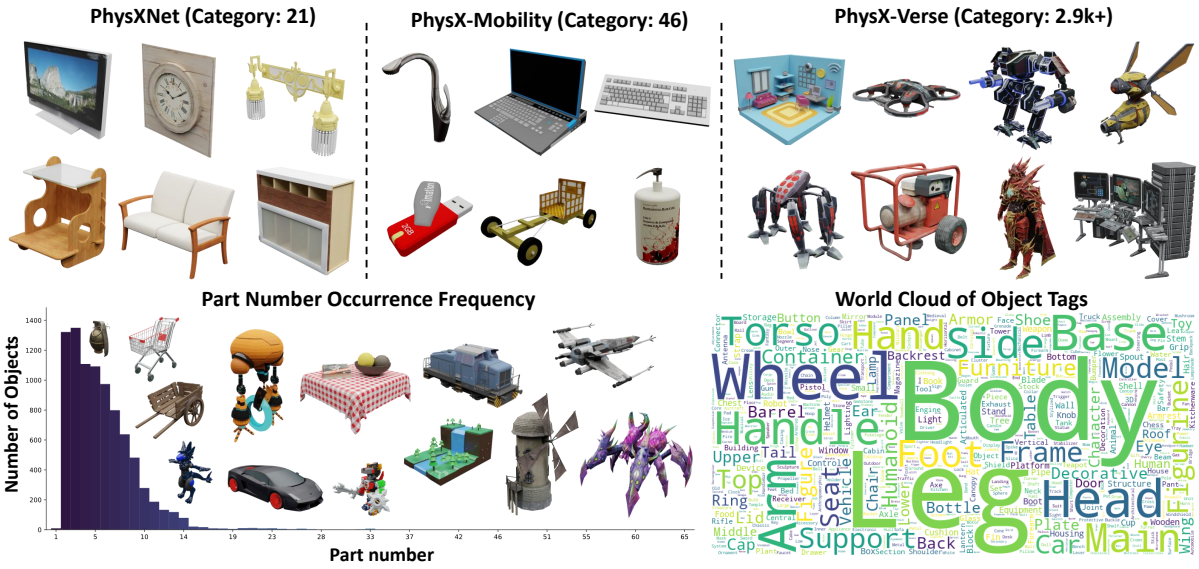

Figure 4: Statistics and distribution of PhysXVerse. Compared to existing simulation-ready physical datasets, PhysXVerse exhibits substantially broader category coverage, including cars, buildings, human models, toys, and robots. The distribution of part counts follows a long-tailed pattern, and the word cloud further illustrates the diversity of semantic categories.

efficiently.

Different from standard 2D RLE, however, 3D structures naturally contain strong spatial redundancy across neighboring slices, especially for smooth or repeated geometric regions. To exploit this property, we further introduce the concept of template layers. Instead of independently encoding every slice, our method allows multiple slices to share a common structural template, while only storing their relative variations or residual differences. By reusing structural patterns across layers, our template-based formulation substantially reduces token redundancy and sequence length while preserving detailed geometric information. Moreover, this design maintains explicit geometric structures throughout the generation process, making it more robust to autoregressive prediction errors and more suitable for high-resolution structure modeling.

As a result, our template-based RLE representation achieves significantly stronger compression efficiency and geometric fidelity compared with conventional 2D RLE and existing text-based explicit representations. We further compare our representation with prior methods in Fig. 3a. The qualitative results demonstrate that, compared with the baseline using text-based voxel indices, PhysX-Omni produces substantially more detailed geometric structures and achieves better alignment with physical and kinematic attributes. In particular, our representation enables the model to maintain structural consistency in complex articulated objects while preserving fine-grained geometry. Additional quantitative and qualitative comparisons are provided in the experimental section.

3.2 PhysXVerse Datasets

To alleviate the limitation of data scarcity, we construct the first general simulation-ready physical 3D dataset, PhysXVerse. To obtain high-quality simulation-ready assets, we leverage the human-verified segmentation annotations provided by PartVerse [23]. For reliable physical properties, we further adopt the human-in-the-loop annotation pipeline introduced in [20]. Specifically, we first preprocess the original dataset by filtering invalid samples and merging excessively small or noisy parts to improve structural consistency. We then render multi-view images of each 3D asset and employ a powerful VLM (GPT) to generate preliminary physical annotations, including absolute scale, affordance, material, functional descriptions, and kinematic information. These automatically generated annotations are subsequently verified and refined by human annotators to ensure both physical plausibility and annotation quality.

As a result, PhysXVerse contains more than 8.7K high-quality simulation-ready 3D assets spanning

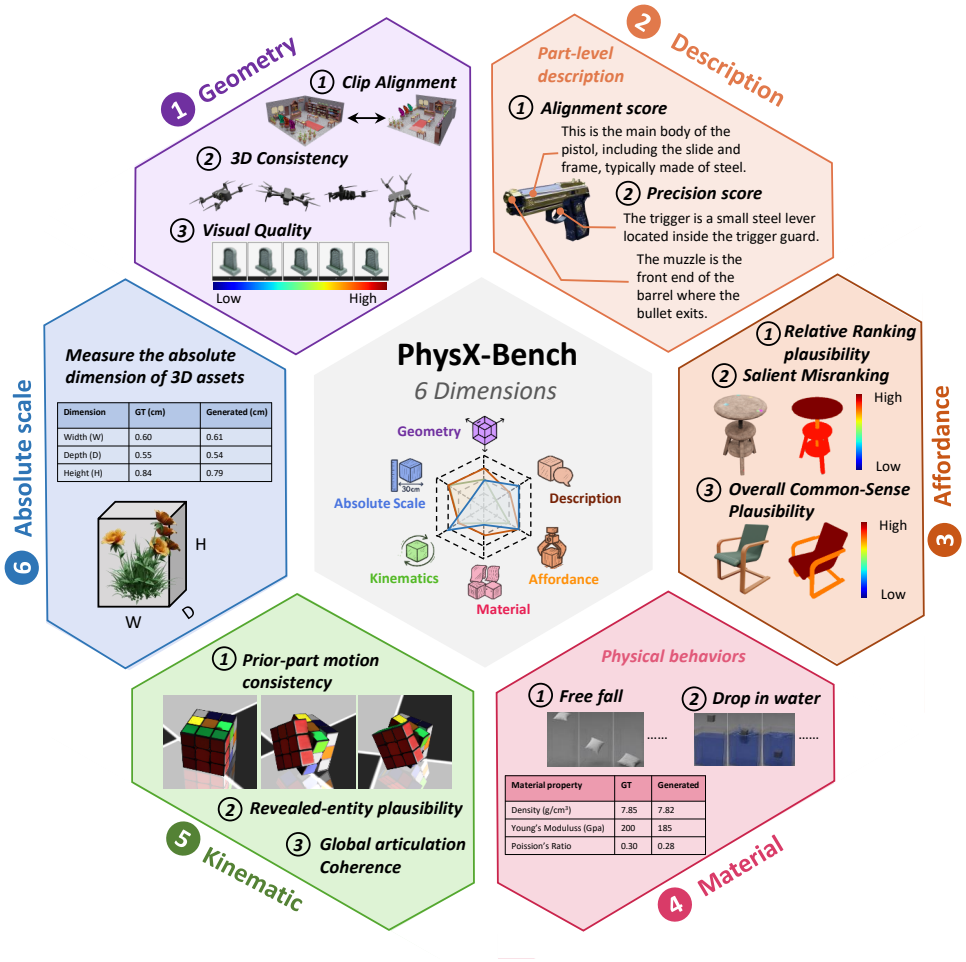

Figure 5: Overview of PhysX-Bench. It consists of six key dimensions for comprehensively evaluating 3D structure, appearance, fundamental physical attributes, and understanding.

over 2.9K categories, covering a wide range of object types, such as indoor furniture, unmanned aerial vehicles, robots, vehicles, and large-scale scene components. Compared with existing simulation-ready datasets, PhysXVerse exhibits substantially richer category diversity and more comprehensive physical annotations, as illustrated in Fig. 4. In addition, we analyze the structural complexity of the dataset through the distribution of part counts. The number of parts ranges from 1 to 65, demonstrating that PhysXVerse covers objects from simple rigid structures to highly complex articulated systems. Such large diversity in both category coverage and structural complexity provides a strong foundation for training and evaluating general simulation-ready physical 3D generation models.

3.3 Evaluation Dimension of PhysX-Bench

To guarantee the reproducibility and robustness of the benchmark, we adopt the open-source VLM (Qwen3.5-122B-A10B) to evaluate the generated physical attributes. Moreover, to reduce the difficulty of understanding complex 3D structures and physical properties, we use rendered images or videos as inputs for evaluation rather than directly feeding physical attributes. Our benchmark evaluates six key dimensions: geometry for evaluating 3D structure and appearance, absolute scale for evaluating physical dimensions, affordance for evaluating human–object interaction priors, description for evaluating semantic understanding, material for evaluating mechanical properties, and kinematics for evaluating motion behaviors shown in Fig 5. Specifically, we define three sub-attributes for geometry: *i.e.*, 1) CLIP score to measure the alignment between the generated results and the conditioning image; 2) 3D consistency to assess structural consistency across multi-view renderings; and 3) visual quality to evaluate the appearance quality. To obtain accurate visual quality assessments, we design a reference grading table

with five levels ranging from very poor to excellent.

For description evaluation, we render part-level masks on the generated 3D object and use the VLM to evaluate whether the masked regions semantically match the human-annotated reference descriptions from the condition image. This assesses whether the evaluated generation method preserves and grounds part-level semantics from the condition image in the generated 3D asset. Since affordance may involve multiple plausible outcomes depending on different functionalities, our evaluation is grounded in human common sense and considers both local and global plausibility, including the relative ranking plausibility and salient misranking of typical parts, as well as the overall rationality of the predicted affordances. Predictions that are more consistent with human common sense will receive higher scores. For absolute scale, we compare the maximum generated object dimension with the VLM-estimated maximum real-world dimension and convert the symmetric percentage error into a scale plausibility score.

For the material dimension, we explore evaluating physical properties by rendering the generated assets into different types of simulation videos, mainly including free-fall and water-drop scenarios. Specifically, the free-fall simulation, particularly the behavior upon ground contact, can reflect properties such as Young's modulus and Poisson's ratio; while the water-drop simulation is mainly used to evaluate density. We believe that evaluating materials through such visualized physical behaviors enables a more intuitive protocol that better aligns with human perception and judgment. For kinematics, we follow the principle that assets with more reasonable and physically plausible motions should receive higher scores. Specifically, we first render the generated assets into motion videos and then infer potential motions from the conditioning image. For visible parts, we define a *prior-part motion consistency* metric to evaluate whether the predicted motions align with the expected behaviors of observed components. For parts that are not visible due to the single-view limitation of the conditioning image but become observable in the rendered motion video, we introduce a *revealed-entity plausibility* metric to assess whether their revealed motions are physically and semantically plausible. Finally, we define a *global articulation coherence* metric to measure the overall consistency and plausibility of the complete motion dynamics. The final kinematics score is computed as a weighted average of the *prior-part motion consistency*, *revealed-entity plausibility*, and *global articulation coherence* scores.

4 Experiments

In this section, we present experimental results on both conventional evaluation metrics and our proposed benchmark, PhysX-Bench. In addition, we report the human alignment evaluation results and conduct comprehensive ablation studies to analyze the effectiveness of different components in our framework. Finally, we further demonstrate the potential applications of PhysX-Omni in downstream simulation-ready scene generation and robotic policy learning tasks.

4.1 Implementation details

We adopt Alibaba Cloud Qwen2.5-VL-7B-Instruct as our VLM backbone [49]. The model is trained for 5 epochs on 64 NVIDIA A100 GPUs over approximately 14 days, using a peak learning rate of 2×10^{-5}, a cosine learning rate decay schedule with a warmup ratio of 0.03, and an effective batch size of 128. To support the generation of high-resolution simulation-ready structures and long-context physical descriptions, we set the maximum sequence length to 16,384 tokens. For the decoding stage, we employ TRELLIS [1] to transform the generated voxel representations into high-quality 3D meshes. Benefiting from our explicit geometry representation, the decoder can directly reconstruct detailed structures without requiring additional mesh segmentation or topology refinement modules, thereby improving both robustness and geometric fidelity.

4.2 Datasets

For training, we combine simulation-ready assets from PhysXNet [20], PhysX-Mobility [21], and our newly constructed dataset PhysXVerse, resulting in a large-scale corpus containing more than 42K

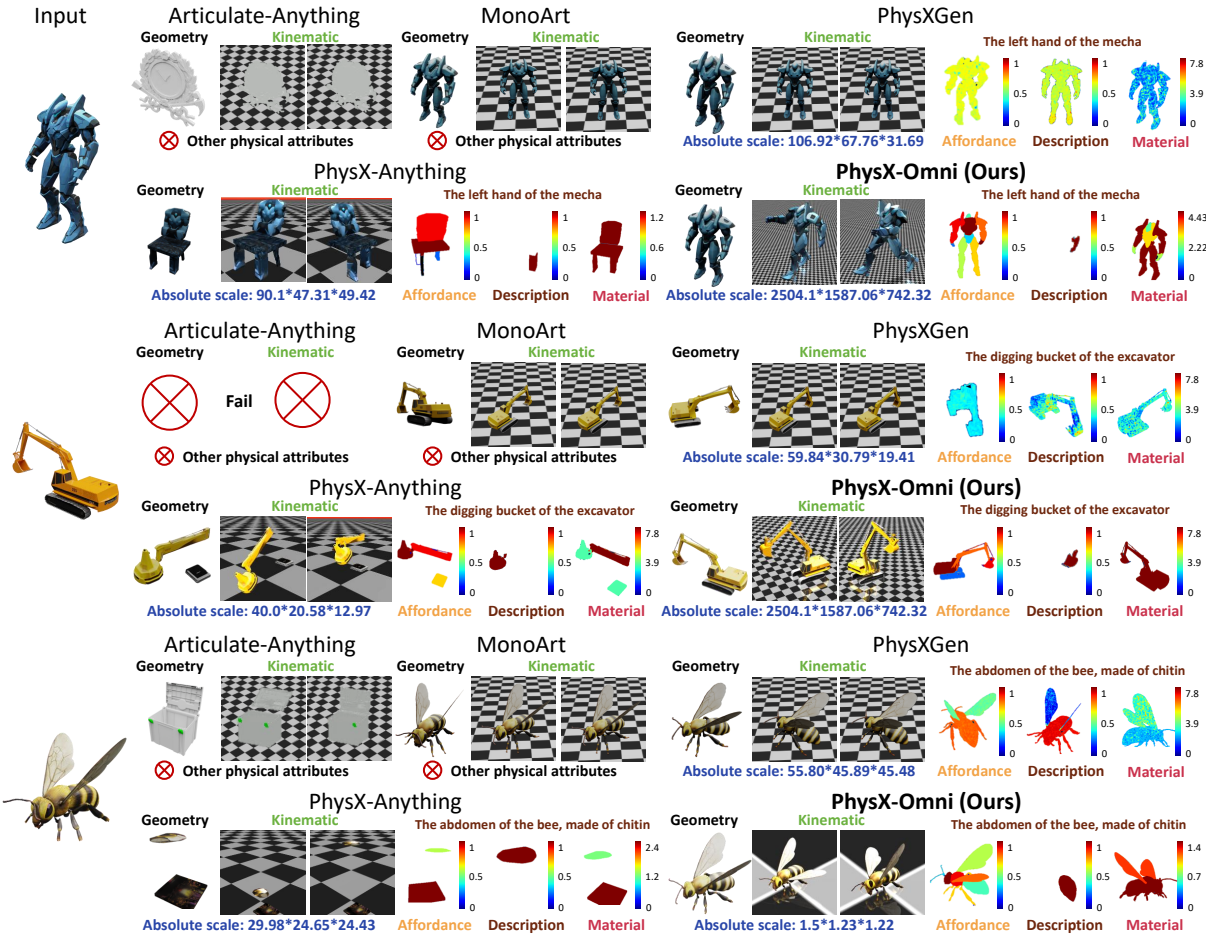

Figure 6: Qualitative results. Compared with existing generative methods, our PhysX-Omni demonstrates impressive performance in generating complex geometries and rich physical attributes.

simulation-ready physical 3D assets spanning diverse indoor and outdoor categories. The dataset covers rigid, articulated, and deformable objects with rich geometric structures and physical attributes. To improve view consistency and enhance the robustness of visual understanding, we render 25 images for each object from different viewpoints as conditioning inputs during training. This multi-view training strategy enables PhysX-Omni to better capture the correspondence between visual appearance, geometric structure, and physical properties, leading to stronger generalization performance on complex real-world objects.

4.3 Conventional evaluation metrics

In our experiments, we evaluate the generated simulation-ready 3D assets using both conventional geometric metrics and physical attribute metrics to comprehensively assess visual fidelity, structural quality, and physical correctness. For geometry evaluation, we adopt Peak Signal-to-Noise Ratio (PSNR) to measure rendered appearance quality, and Chamfer Distance (CD) together with F-score to evaluate the accuracy of reconstructed 3D geometry. To ensure robustness and reduce viewpoint bias, we render both the generated assets and the ground-truth assets from 30 different viewpoints and compute the averaged evaluation results across all rendered views. Beyond geometric quality, we further evaluate the generated physical attributes following the protocol of [21]. Specifically, for absolute scale evaluation, we compute the Mean Squared Error (MSE) between the predicted and ground-truth object scales. For material, affordance, and description evaluation, we adopt heatmap-based PSNR metrics to measure the similarity between the predicted physical attribute distributions and the corresponding ground-truth annotations. These metrics provide a more robust evaluation of semantic and physical consistency in generated assets.

Table 1: Quantitative comparison with other methods on conventional metrics. Note that Chamfer Distance (CD) is reported in units of $\times 10^{-3}$, and F-score (FS) is reported in units of $\times 10^{-2}$ under a distance threshold of 0.05. The results clearly demonstrate the superior generative performance of our method in both geometry and physical attribute generation.

Dataset	Methods	Geometry			Physical Attributes				
		PSNR ↑	CD ↓	F-score ↑	Absolute scale ↓	Material ↑	Affordance ↑	Kinematic ↑	Description ↑
PhysXVerse	Articulate-Anything [11]	14.03	48.77	46.44	–	–	–	0.2952	–
	MonoArt [13]	19.68	7.03	85.27	–	–	–	0.3805	–
	PhysXGen [20]	19.41	15.19	83.56	309.31	16.51	9.40	0.3494	11.84
	PhysX-Anything [21]	15.97	37.06	40.46	298.19	15.65	10.50	0.4191	21.38
	PhysX-Omni (Ours)	**21.52**	**2.95**	**91.28**	**2.79**	**27.23**	**21.47**	**0.9185**	**31.05**
PhysX-Mobility	Articulate-Anything [11]	15.02	16.09	66.95	–	–	–	0.6396	–
	MonoArt [13]	16.46	6.35	87.41	–	–	–	0.4351	–
	PhysXGen [20]	15.75	35.32	79.62	46.58	16.02	8.73	0.3884	11.60
	PhysX-Anything [21]	16.57	23.13	**89.51**	22.58	22.58	16.29	0.7852	26.28
	PhysX-Omni (Ours)	**18.38**	**4.70**	88.50	**2.78**	**24.09**	**16.58**	**0.8603**	**28.40**

Table 2: Quantitative comparison with other methods on PhysX-Bench. The results validate the strong generalization capability of our method on both real-world and complex synthetic images, achieving significant improvements over all competing methods.

Methods	Geometry			Physical Attributes				
	CLIP ↑	3D Consistency ↑	Visual Quality ↑	Absolute scale ↑	Material ↑	Affordance ↑	Kinematic ↑	Description ↑
Articulate-Anything [11]	0.554	55.27	88.46	–	–	–	71.25	–
MonoArt [13]	**0.835**	**82.56**	**96.20**	–	–	–	68.32	–
PhysXGen [20]	0.803	73.50	85.93	24.21	–	66.07	69.17	22.24
PhysX-Anything [21]	0.547	52.71	70.81	50.20	44.70	59.96	65.99	26.89
PhysX-Omni (Ours)	0.767	64.48	90.0	**64.26**	**59.89**	**70.57**	**80.72**	**39.02**

For kinematic evaluation, we measure the MSE between the predicted and ground-truth articulation parameters, including joint axis positions, joint directions, joint types, and motion limits. This evaluation particularly assesses whether the generated assets can accurately capture physically plausible articulated behaviors required for downstream simulation and robotic interaction. By jointly evaluating geometry, physical attributes, and articulation properties, our evaluation protocol provides a comprehensive assessment of simulation-ready physical 3D generation quality.

4.4 Evaluations with conventional metrics

We compare PhysX-Omni with several recent simulation-ready 3D generation methods, including PhysXGen [20], Articulate-Anything [11], MonoArt [13], and PhysX-Anything [21]. Following the evaluation protocols of PhysXGen [20] and MonoArt [13], we conduct experiments on both PhysXVerse and PhysX-Mobility datasets using conventional geometric metrics and physical attribute evaluations.

As shown in Table. 1, PhysX-Omni consistently achieves the best performance across nearly all evaluation metrics on both datasets, demonstrating the effectiveness of our unified simulation-ready generation framework. In terms of geometric quality, PhysX-Omni significantly outperforms previous methods on PSNR, Chamfer Distance (CD), and F-score. On PhysXVerse, our method achieves a PSNR of 21.52, CD of 2.95, and F-score of 91.28, substantially surpassing the previous best results. These improvements indicate that our method can generate more accurate and structurally consistent 3D geometry with finer local details. The strong geometric performance mainly benefits from our tailored template-based geometry representation. Unlike previous methods that rely on text-based voxel indices or additional segmentation modules, our representation directly models high-resolution structures in an explicit and compact manner. This design effectively reduces segmentation-induced artifacts and improves the consistency between neighboring geometric regions. As a result, PhysX-Omni generates cleaner object boundaries, more detailed local structures, and more coherent articulated components, especially for objects with complex topologies and fine-grained geometry.

Beyond geometric generation, PhysX-Omni also demonstrates substantial improvements in physical attribute prediction. In particular, our method achieves remarkably lower absolute scale errors compared with previous approaches. On PhysXVerse, the absolute scale error is reduced from 309.31 in PhysXGen and 298.19 in PhysX-Anything to only 2.79 in PhysX-Omni. A similar improvement is observed on

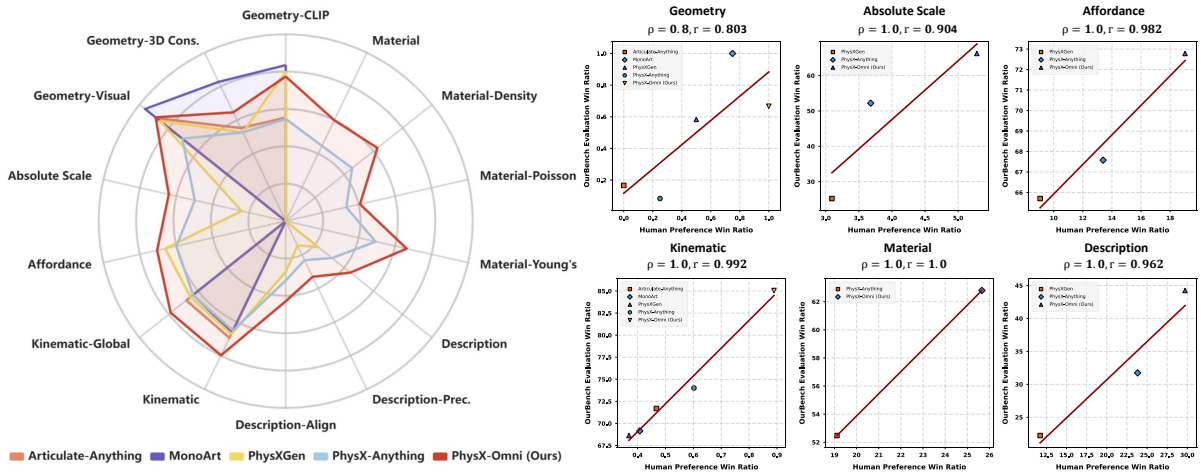

Figure 7: Left: Comparison of our PhysX-Omni with other methods. It validate the impressive overall performance of PhysX-Omni. **Right: Human alignment validation of PhysX-Bench.** Our experiments show that the our PhysX-Bench across all dimensions closely align with human annotations.

PhysX-Mobility, where the error decreases to 2.78. These results demonstrate that our framework possesses significantly stronger understanding of real-world object dimensions and physical priors, which is essential for downstream simulation and robotic interaction tasks. For material, affordance, and description evaluations, PhysX-Omni also consistently achieves the best performance across both datasets. On PhysXVerse, our method improves the material score from 15.65 to 27.23 and the affordance score from 10.50 to 21.47 compared with PhysX-Anything. Similar trends can be observed on PhysX-Mobility. These improvements indicate that PhysX-Omni can better capture semantic functionality and physical properties of objects, producing more realistic and physically plausible simulation-ready assets. Notably, the most significant gains are achieved in kinematic evaluation. On PhysXVerse, PhysX-Omni reaches a kinematic score of 0.9185, greatly outperforming previous methods such as PhysX-Anything (0.4191) and MonoArt (0.3805). On PhysX-Mobility, our method similarly achieves a strong kinematic score of 0.8603. These results validate that our framework can accurately infer articulation structures, joint types, and motion constraints, enabling the generation of articulated assets with physically consistent behaviors.

Overall, the quantitative results demonstrate that PhysX-Omni achieves superior performance in both geometry and physical reasoning. Benefiting from the combination of our VLM-based global-to-local framework and the proposed high-resolution geometry representation, PhysX-Omni produces simulation-ready assets with higher visual fidelity, stronger physical consistency, and more accurate articulation modeling. These results collectively validate the superiority, robustness, and generalizability of our unified framework for simulation-ready physical 3D generation across diverse object categories and physical settings.

4.5 Evaluations on PhysX-Bench

To comprehensively evaluate the generalization ability of different methods in real-world scenarios, we further conduct evaluations on our newly proposed benchmark, PhysX-Bench. More details of the benchmark construction and evaluation metrics are provided in the supplementary material. As mentioned previously, the conditioning images in PhysX-Bench are collected from both real-world photographs and rendered images of 3D assets, covering a wide range of common object categories and challenging in-the-wild cases. Compared with conventional benchmarks that rely on ground-truth annotations, PhysX-Bench emphasizes ground-truth-free evaluation across multiple dimensions, including geometry, absolute scale, material, affordance, kinematics, and semantic description.

The quantitative results in Table 2 strongly demonstrate the superior overall performance of PhysX-Omni compared with existing approaches. In particular, our method achieves the best results on most physical attributes, including absolute scale, material, affordance, kinematics, and description. Specif-

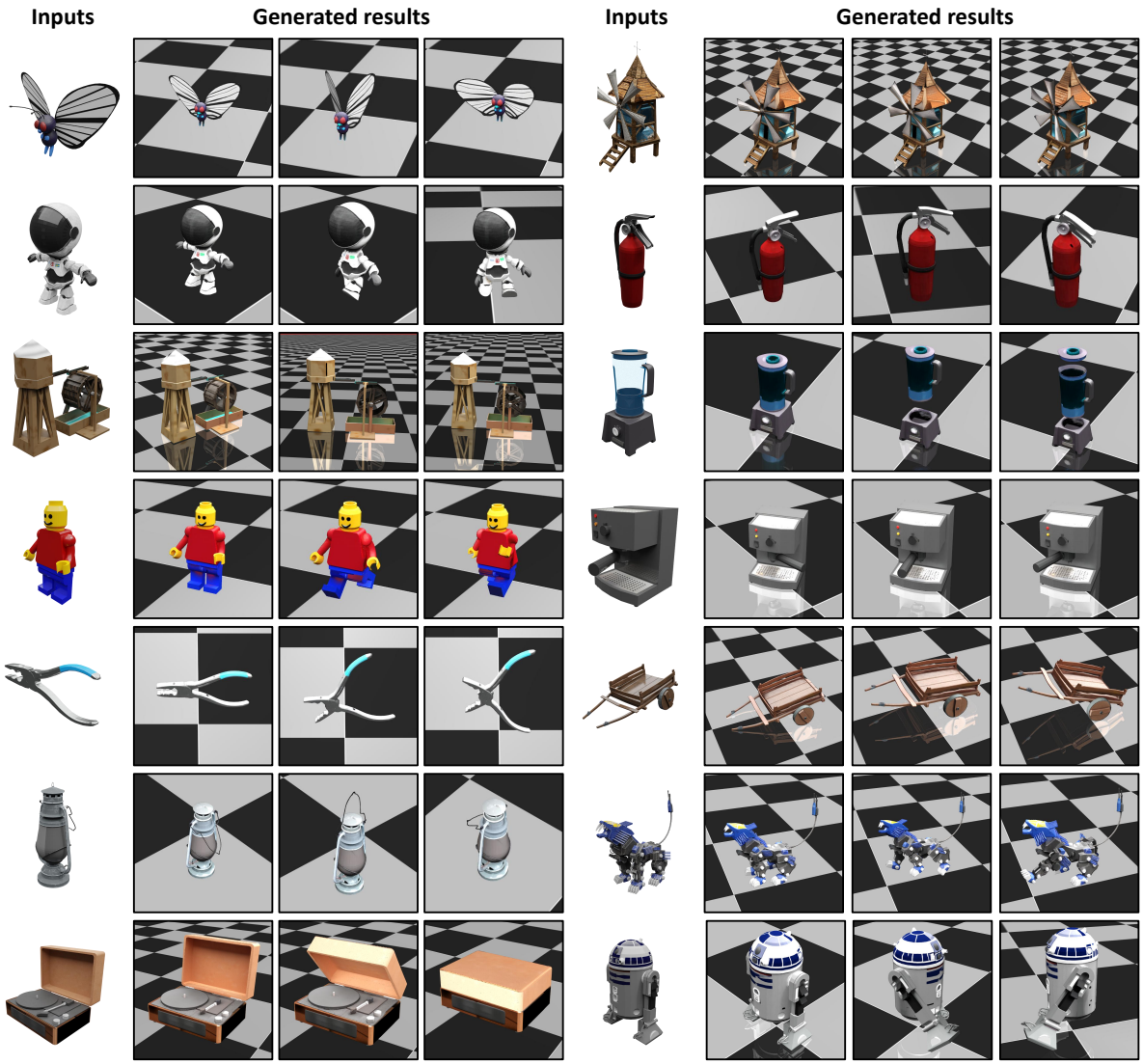

Figure 8: More qualitative results of PhysX-Omni. Additional results further demonstrate the robust generative performance of our method in complex scenarios.

ically, PhysX-Omni achieves a kinematic score of 80.72, significantly outperforming PhysX-Anything (65.99), PhysXGen (69.17), MonoArt (68.32), and Articulate-Anything (71.25). Similar improvements can also be observed for affordance understanding and description quality, validating the strong physical reasoning and semantic understanding capabilities of our framework.

Benefiting from the explicitly encoded high-resolution 3D structures, PhysX-Omni can better model the intrinsic interdependency between geometry, articulation, and physical attributes. Unlike prior methods that heavily rely on segmentation-based intermediate representations, our framework directly models explicit 3D structures in a unified manner, thereby significantly improving structural coherence and articulation consistency. This advantage is particularly important for complex articulated and deformable objects, where geometric details and kinematic properties are highly coupled. Although MonoArt achieves slightly better performance on several geometry-related metrics, including CLIP similarity, 3D consistency, and visual quality, this advantage mainly arises from its complete reliance on the pre-trained TRELLIS geometry generation pipeline. As a result, MonoArt lacks explicit understanding of part-level motion and physical interactions. Consequently, it exhibits limited capability in modeling physical properties and articulated behaviors, leading to notably inferior performance on simulation-oriented attributes such as kinematics, affordance, and absolute scale. In comparison, PhysX-Omni achieves a much more balanced and robust performance across both geometry and physical reasoning tasks shown in Fig. 7. The results demonstrate that our method not only preserves strong geometric quality, but

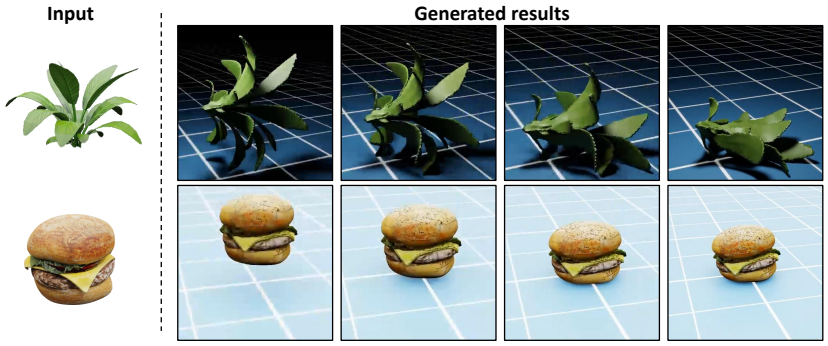

Figure 9: Visualization of the generated deformable objects. It illustrates the realistic deformation behavior of our generated deformable assets during free fall under physical simulation.

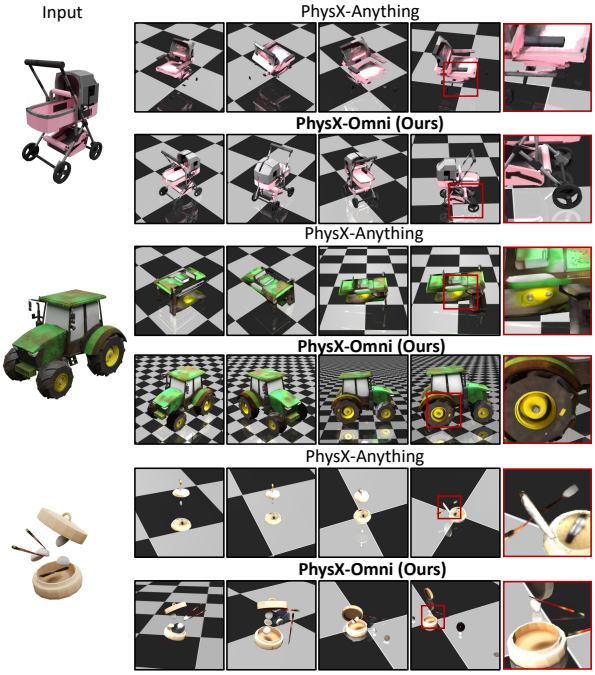

Figure 10: Visualization of model using different geometry representations. It strongly demonstrates that, by introducing an efficient geometry representation, PhysX-Omni achieves superior performance in generating complex structures compared with our baseline.

also substantially improves the generation of physically plausible simulation-ready assets. In particular, our explicit geometry representation enables the model to maintain detailed structures while avoiding segmentation-induced ambiguities and artifacts, resulting in more coherent articulated motions and more reliable physical attributes.

To provide more intuitive comparisons, we further visualize the generated results of different methods in Fig. 6. The qualitative results show that PhysX-Omni demonstrates remarkable robustness in generating complex structures and challenging articulated objects. Compared with the baseline PhysX-Anything, our method produces higher-quality simulation-ready assets with more accurate geometry, more plausible material and affordance predictions, and significantly more coherent articulated behaviors. Furthermore, additional visualization results in Fig. 8 and Fig. 9 demonstrate the capability of PhysX-Omni in generating diverse simulation-ready scenes involving rigid, deformable, and articulated objects. By directly modeling explicit 3D structures without relying on additional segmentation modules, PhysX-Omni effectively reduces structural ambiguities and segmentation-induced errors, further validating the effectiveness of our tailored geometry representation and unified simulation-ready generation framework.

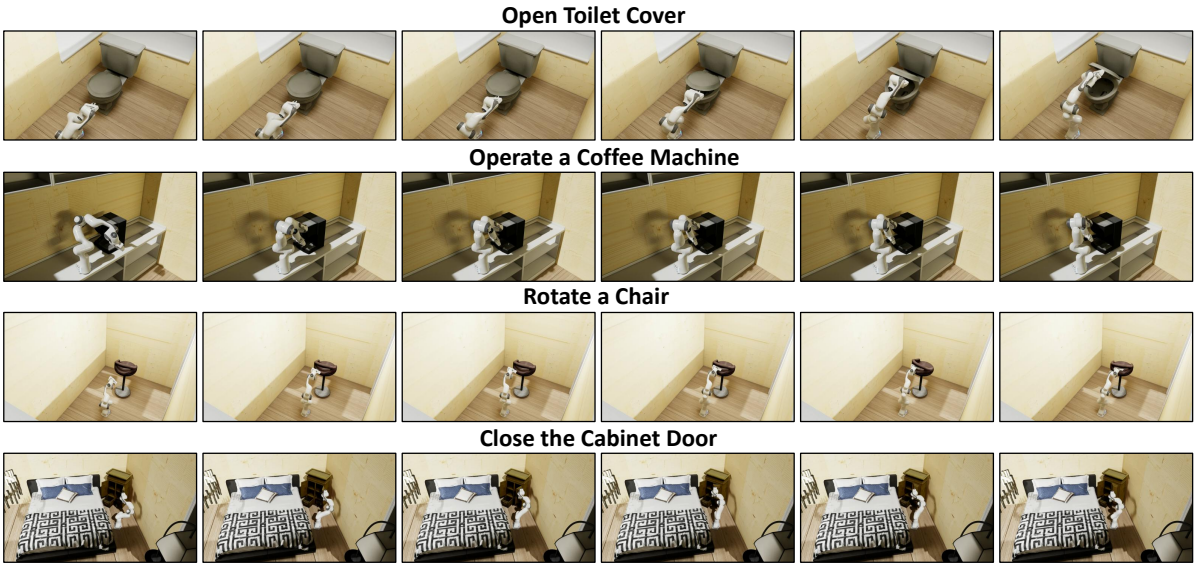

Figure 11: Robot Manipulation on our Generated sim-ready 3D assets. The results demonstrate that our generated simulation-ready assets exhibit highly physically plausible behaviors and accurate geometric structures across diverse tasks, thereby opening a new direction for robotic policy learning.

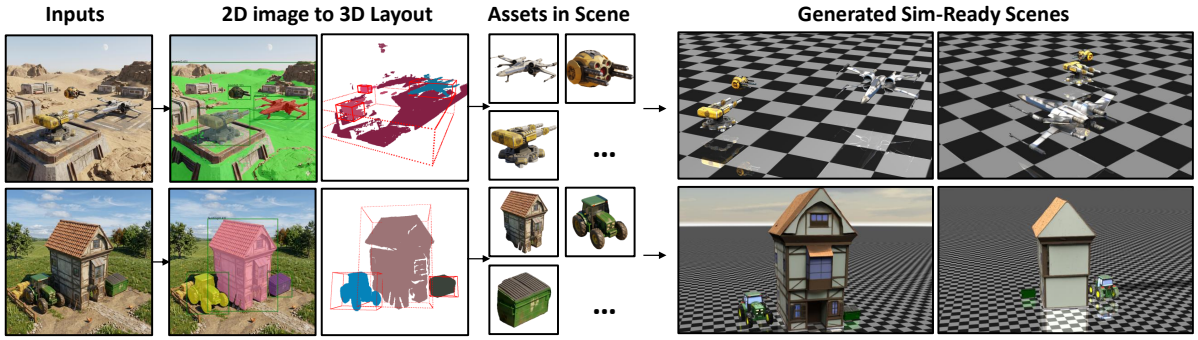

Figure 12: Applications of our PhysX-Omni. We explore the potential applications of PhysX-Omni in sim-ready scene generation.

4.6 Validating human alignment of PhysX-Bench

To validate that PhysX-Bench can effectively reflect human perception and evaluation preferences, we further study the correlation between the automatic evaluation results produced by PhysX-Bench and human annotations. Specifically, following prior benchmark protocols, we measure the alignment between automatic evaluation scores and human preference scores using Spearman's rank correlation coefficient. A higher Spearman correlation coefficient indicates stronger consistency between the benchmark evaluation and human judgment. As shown in Fig. 7, PhysX-Bench demonstrates consistently strong correlations with human annotations across all major evaluation dimensions, including geometry, absolute scale, affordance, kinematics, material, and semantic description. In particular, several physical attributes achieve superior rank consistency with human preferences. For example, absolute scale, affordance, material, and description all achieve a Spearman coefficient of $\rho = 1.0$, while kinematic evaluation reaches $\rho = 1.0$ with an exceptionally high Pearson correlation coefficient of $r = 0.992$. These results indicate that the rankings produced by PhysX-Bench are highly aligned with human evaluation outcomes. Moreover, even for geometry evaluation, which is generally more challenging due to the diversity of visual appearances and structural details, PhysX-Bench still achieves a strong correlation with human preferences ($\rho = 0.8$, $r = 0.803$). This result further demonstrates that our benchmark can robustly evaluate not only physical attributes but also geometric quality in complex real-world scenarios.

Overall, the strong correlations across all evaluation dimensions validate the reliability, robustness, and effectiveness of PhysX-Bench. These results demonstrate that our benchmark can serve as a trust-

worthy automatic evaluation framework for simulation-ready physical 3D generation, providing evaluation results that closely match human perception and judgment.

4.7 Ablation Studies

To validate the effectiveness of our tailored geometry representation, we compare PhysX-Omni with a baseline that directly employs text-based voxel indices to model 3D structures. The quantitative results on both conventional metrics and PhysX-Bench, reported in Table 1 and Table 2, consistently demonstrate the substantial improvements brought by our proposed template-based geometry representation. In particular, our method achieves significantly better performance on kinematic and absolute scale, validating the effectiveness of explicitly modeling high-resolution structures in a compact and generation-friendly manner.

Furthermore, the qualitative comparisons shown in Fig. 10 provide more intuitive evidence of the advantages of our representation. Compared with the baseline PhysX-Anything, which relies on text-based voxel indices and additional segmentation processes, PhysX-Omni produces substantially more detailed and structurally coherent simulation-ready assets. As illustrated in the highlighted regions, the baseline method frequently suffers from structural ambiguities, incomplete local geometry, and inconsistent articulated components, especially for objects with complex topologies and fine-grained structures, such as strollers and tractors. In contrast, by directly modeling explicit 3D geometry and eliminating the additional segmentation stage during generation, PhysX-Omni effectively reduces segmentation-induced artifacts and error accumulation. This design enables our method to better preserve local geometric continuity and part-level structural consistency, leading to sharper details, more accurate articulated structures, and more physically plausible object layouts. For example, PhysX-Omni can generate more accurate wheel structures, cleaner articulated connections, and more stable local geometries in highly complex regions where the baseline often fails.

Moreover, the improvements are particularly evident for articulated objects and objects involving strong part interactions. Benefiting from the explicit structural representation, PhysX-Omni can better capture the intrinsic relationships between geometry and kinematics, thereby improving both structural reasoning and motion consistency. These results collectively demonstrate that our tailored geometry representation significantly enhances the robustness, fidelity, and generalization ability of simulation-ready physical 3D generation.

4.8 Application: Robotic Policy Learning in Simulation

As shown in Fig. 11, we further investigate whether the generated assets can be effectively utilized in real simulation environments and downstream robotic tasks. To this end, we directly deploy the generated simulation-ready 3D assets into a physics simulator for robotic interaction and policy learning. Specifically, the generated assets are imported together with their geometric structures, physical properties, and articulated parameters, enabling the simulator to perform physically grounded interactions without additional manual processing. The experimental results demonstrate that our generated assets maintain reliable geometric accuracy, physically plausible material properties, and coherent articulated behaviors under dynamic interactions. Even in challenging manipulation scenarios involving articulated motion and object contact, the generated assets remain structurally stable and physically consistent. These results suggest that PhysX-Omni not only produces visually realistic 3D assets, but also generates physically functional representations that can be seamlessly integrated into simulation pipelines for downstream robotics and embodied AI applications. Furthermore, the ability to automatically generate simulation-ready assets from in-the-wild images significantly reduces the cost of manual asset construction for robotic training environments.

4.9 Application: Sim-Ready Scene Generation

In addition, we further explore the potential of PhysX-Omni for scene-level simulation-ready generation. Specifically, we first employ image-to-depth estimation methods [50] together with 2D segmentation approaches [51] to reconstruct an initial 3D scene layout from input images. Based on the estimated depth, segmentation masks, and scene geometry, we obtain coarse object placements and spatial relationships within the environment. We then integrate the reconstructed 3D layout with the simulation-ready assets generated by PhysX-Omni to automatically build physically plausible simulation-ready scenes, as illustrated in Fig. 12. Benefiting from the explicit geometric structures and physical attributes generated by our framework, the inserted assets can maintain consistent scales. Moreover, since our framework supports rigid, deformable, and articulated objects in a unified manner, it enables the construction of significantly more diverse and realistic simulation environments compared with previous approaches.

These results demonstrate that PhysX-Omni not only supports high-quality simulation-ready asset generation, but also provides a promising foundation for scalable scene-level simulation construction, embodied AI training, robotic policy learning, and future physically grounded world generation applications.

5 Conclusion

In this paper, we introduce **PhysX-Omni**, a unified framework for simulation-ready physical 3D generation across diverse asset types, including rigid, deformable, and articulated objects. By proposing a tailored geometry representation for vision–language models, PhysX-Omni directly models detailed 3D structures without introducing additional special tokens or relying on segmentation modules, thereby significantly improving generation quality and robustness. To alleviate the limitation of data scarcity, we further construct the first general simulation-ready physical 3D dataset, PhysXVerse, containing over 8.7K high-quality assets with rich physical annotations. Moreover, to evaluate simulation-ready 3D generation in real-world scenarios, we propose a new benchmark, PhysX-Bench, which performs ground-truth-free evaluation across six key dimensions, including geometry, absolute scale, affordance, material, kinematics, and semantic description. Comprehensive experiments on both PhysX-Bench and conventional evaluation metrics demonstrate the superior performance and strong generalization ability of PhysX-Omni. Furthermore, additional studies validate the potential of our framework in downstream applications such as simulation-ready scene generation and robotic policy learning, highlighting the feasibility of directly deploying generated assets into embodied AI and robotic simulation environments.

Limitation and future work. Despite the strong performance of PhysX-Omni, several limitations remain. In particular, geometric quality can still be improved for highly complex structures and fine-grained details. Since our framework emphasizes unified physical understanding and simulation-ready generation rather than appearance-oriented geometry pre-training, it may underperform on certain appearance-focused geometric metrics. In the future, we plan to leverage larger-scale 3D geometry datasets and stronger appearance supervision to further enhance geometric fidelity while maintaining physical consistency.

5.1 Acknowledgments

This research is supported by cash and in-kind funding from NTU S-Lab and industry partner(s). This study is also supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012, MOE-T2EP20223-0002).

References

[1] 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.

[2] Fangzhou Hong, Jiaxiang Tang, Ziang Cao, Min Shi, Tong Wu, Zhaoxi Chen, Shuai Yang, Tengfei Wang, Liang Pan, Dahua Lin, et al. 3dtopia: Large text-to-3d generation model with hybrid diffusion priors. *arXiv preprint arXiv:2403.02234*, 2024.

[3] Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. *arXiv preprint arXiv:2409.12957*, 2024.

[4] Shuangkang Fang, I Shen, Yufeng Wang, Yi-Hsuan Tsai, Yi Yang, Shuchang Zhou, Wenrui Ding, Takeo Igarashi, Ming-Hsuan Yang, et al. Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14061–14072, 2025.

[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.

[6] Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025.

[7] Longwen Zhang, Qixuan Zhang, Haoran Jiang, Yinuo Bai, Wei Yang, Lan Xu, and Jingyi Yu. Bang: Dividing 3d assets via generative exploded dynamics. *ACM Transactions on Graphics (TOG)*, 44(4):1–21, 2025.

[8] 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. *arXiv preprint arXiv:2507.06165*, 2025.

[9] 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.

[10] Jiayi Liu, Denys Iliash, Angel X Chang, Manolis Savva, and Ali Mahdavi-Amiri. Singapo: Single image controlled generation of articulated parts in objects. *arXiv preprint arXiv:2410.16499*, 2024.

[11] 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.

[12] Ruijie Lu, Yu Liu, Jiaxiang Tang, Junfeng Ni, Yuxiang Wang, Diwen Wan, Gang Zeng, Yixin Chen, and Siyuan Huang. Dreamart: Generating interactable articulated objects from a single image. *arXiv preprint arXiv:2507.05763*, 2025.

[13] 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*, 2026.

[14] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pages 388–406, 2024.

[15] Minghao Guo, Bohan Wang, Pingchuan Ma, Tianyuan Zhang, Crystal Owens, Chuang Gan, Josh Tenenbaum, Kaiming He, and Wojciech Matusik. Physically compatible 3d object modeling from a single image. *Advances in Neural Information Processing Systems*, 37:119260–119282, 2024.

[16] Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6178–6189, 2025.

[17] Long Le, Ryan Lucas, Chen Wang, Chuhao Chen, Dinesh Jayaraman, Eric Eaton, and Lingjie Liu. Pixie: Fast and generalizable supervised learning of 3d physics from pixels. *arXiv preprint arXiv:2508.17437*, 2025.

[18] Chuhao Chen, Zhiyang Dou, Chen Wang, Yiming Huang, Anjun Chen, Qiao Feng, Jiatao Gu, and Lingjie Liu. Vid2sim: Generalizable, video-based reconstruction of appearance, geometry and physics for mesh-free simulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26545–26555, 2025.

[19] Hanxiao Jiang, Hao-Yu Hsu, Kaifeng Zhang, Hsin-Ni Yu, Shenlong Wang, and Yunzhu Li. Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos. *arXiv preprint arXiv:2503.17973*, 2025.

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

[21] 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*, 2025.

[22] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4209–4219, 2024.

[23] Shaocong Dong, Lihe Ding, Xiao Chen, Yaokun Li, Yuxin Wang, Yucheng Wang, Qi Wang, Jaehyeok Kim, Chenjian Gao, Zhanpeng Huang, et al. From one to more: Contextual part latents for 3d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8230–8240, 2025.

[24] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022.

[25] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances in neural information processing systems*, 35:31841–31854, 2022.

[26] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[27] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[28] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

[29] Ziang Cao, Fangzhou Hong, Tong Wu, Liang Pan, and Ziwei Liu. Large-vocabulary 3d diffusion model with transformer. *arXiv preprint arXiv:2309.07920*, 2023.

[30] Ziang Cao, Fangzhou Hong, Tong Wu, Liang Pan, and Ziwei Liu. Difftf++: 3d-aware diffusion transformer for large-vocabulary 3d generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[31] Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Collaborative multi-modal coding for high-quality 3d generation. *arXiv preprint arXiv:2508.15228*, 2025.

[32] Yunhan Yang, Yuan-Chen Guo, Yukun Huang, Zi-Xin Zou, Zhipeng Yu, Yangguang Li, Yan-Pei Cao, and Xihui Liu. Holopart: Generative 3d part amodal segmentation. *arXiv preprint arXiv:2504.07943*, 2025.

[33] Runmao Yao, Junsheng Zhou, Zhen Dong, and Yu-Shen Liu. Anchoreddream: Zero-shot 360 {\deg} indoor scene generation from a single view via geometric grounding. *arXiv preprint arXiv:2601.16532*, 2026.

[34] 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. *arXiv preprint arXiv:2506.05573*, 2025.

[35] ByteDance Seed. Seed3d 1.0: From images to high-fidelity simulation-ready 3d assets. 2025.

[36] Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, et al. Meshanything: Artist-created mesh generation with autoregressive transformers. *arXiv preprint arXiv:2406.10163*, 2024.

[37] 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*, pages 19615–19625, 2024.

[38] Zhengyi Wang, Jonathan Lorraine, Yikai Wang, Hang Su, Jun Zhu, Sanja Fidler, and Xiaohui Zeng. Llama-mesh: Unifying 3d mesh generation with language models. *arXiv preprint arXiv:2411.09595*, 2024.

[39] 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. *arXiv preprint arXiv:2502.02590*, 2025.

[40] Chaoyue Song, Jianfeng Zhang, Xiu Li, Fan Yang, Yiwen Chen, Zhongcong Xu, Jun Hao Liew, Xiaoyang Guo, Fayao Liu, Jiashi Feng, et al. Magicarticulate: Make your 3d models articulation-ready. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15998–16007, 2025.

[41] Jiayi Su, Youhe Feng, Zheng Li, Jinhua Song, Yangfan He, Botao Ren, and Botian Xu. Artformer: Controllable generation of diverse 3d articulated objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1894–1904, 2025.

[42] Honghua Chen, Yushi Lan, Yongwei Chen, and Xingang Pan. Artilatent: Realistic articulated 3d object generation via structured latents. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025.

[43] Chuhao Chen, Isabella Liu, Xinyue Wei, Hao Su, and Minghua Liu. Freeart3d: Training-free articulated object generation using 3d diffusion. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–13, 2025.

[44] Abhishek Joshi, Beining Han, Jack Nugent, Max Gonzalez Saez-Diez, Yiming Zuo, Jonathan Liu, Hongyu Wen, Stamatis Alexandropoulos, Karhan Kayan, Anna Calveri, et al. Procedural generation of articulated simulation-ready assets, 2025. 6. *URL https://arxiv. org/abs/2505.10755*, 7.

[45] Jiahui Lei, Congyue Deng, Bokui Shen, Leonidas Guibas, and Kostas Daniilidis. Nap: Neural 3d articulation prior. *arXiv preprint arXiv:2305.16315*, 2023.

[46] Zhe Li, Xiang Bai, Jieyu Zhang, Zhuangzhe Wu, Che Xu, Ying Li, Chengkai Hou, and Shanghang Zhang. Urdf-anything: Constructing articulated objects with 3d multimodal language model. *arXiv preprint arXiv:2511.00940*, 2025.

[47] 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.

[48] Junyi Cao and Evangelos Kalogerakis. Sophy: Generating simulation-ready objects with physical materials. *arXiv preprint arXiv:2504.12684*, 2025.

[49] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[50] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[51] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.