PhysX-Anything: Simulation-Ready Physical 3D Assets from Single Image

Supplementary Material

1. Implementation details

Our training pipeline follows a two-stage scheme: (1) VLM training and (2) controllable flow transformer training. In Stage 1, we fine-tune Qwen2.5-VL-7B-Instruct [1] as our vision–language backbone. The model is trained for 5 epochs on 32 NVIDIA A800 GPUs over the course of 5 days, with a peak learning rate of 2×10^{-5}, a cosine learning rate schedule with a warmup ratio of 0.03, and an effective batch size of 128. The maximum sequence length is set to 8,192 tokens. In Stage 2, we freeze the flow transformer and train the controllable branch using the AdamW optimizer with a learning rate of 1×10^{-4} and a global batch size of 32 on 8 NVIDIA A800 GPUs for 1 day.

2. Datasets.

For training, we collect data from PhysXNet [2] and our new dataset PhysX-Mobility, resulting in over 33K physical 3D assets spanning 47 categories. We split our PhysX-Mobility into 1,636 objects for training and 388 for testing. For each object, we render 25 images as conditioning inputs for both the VLM and the geometry decoder, and we construct the training set as a balanced mixture of samples from PhysXNet and PhysX-Mobility.

3. Evaluation metrics

Following the evaluation protocol in [2], we evaluate generative performance along five physical attributes: **absolute scale**, **material**, **affordance**, **kinematics**, and **descriptors**. Specifically, we quantify absolute scale using Euclidean distance, and evaluate density, affordance, and descriptor images using Peak Signal-to-Noise Ratio (PSNR). To more faithfully evaluate **kinematics**, we further employ a VLM-based evaluator (GPT-5) to score the generated articulation results. In addition, to obtain a more comprehensive evaluation, we consider two settings: (1) evaluations on the test set of PhysX-Mobility; and (2) in-the-wild evaluations including VLM-based scores and user studies. Note that, since GPT-5 is not yet fully reliable for fine-grained physical reasoning, we restrict VLM-based evaluation to geometry and kinematic parameters (articulation), while human studies are used to judge overall physical plausibility on real-world images.

4. Detailed information of PhysX-Anything

As discussed in the main paper, PhysX-Anything adopts a multi-round conversation strategy. It first generates global, high-level information about the object, and then sequentially refines it by producing detailed geometric information for each part. The prompts used for the global and geometric stages are given below.

4.1. Prompt for overall information

```
Please analyze the given image of
an object and output its structured
description in the following format
(voxel grid=32):
Name: <object name>
Category: <object category>
Dimension: <physical dimensions in
cm like 50*40*30>
Parts:
l_<id>: <part name>, <Affordance
rank>, <Material Name>, <Density>,
<Young's Modulus>, <Poisson's
Ratio>, <basic description of the
part>
l_<id>: <part name>, <Affordance
rank>, <Material Name>, <Density>,
<Young's Modulus>, <Poisson's
Ratio>, <basic description of the
part>
...
Group_info:
group_<id>: [l_<id>, ...]
(child); Type: E; Params: N/A
group_<id>: [l_<id>, ...]
(child); Type: A relative to
group_<id>(parent); Params: N/A
group_<id>: [l_<id>, ...]
(child); Type: B relative to
group_<id>(parent); Params:
direction: [x,y,z], slide range
(in voxel grid): [min,max]
group_<id>: [l_<id>, ...]
(child); Type: C relative to
group_<id>(parent); Params:
direction: [x,y,z], axis position
(in voxel grid): [x,y,z], revolute
range (degree): [min,max]
group_<id>: [l_<id>, ...]
(child); Type: D relative to
group_<id>(parent); Params: hinge
position (in voxel grid): [x,y,z]
group_<id>: [l_<id>, ...]
(child); Type: CB relative to
group_<id>(parent); Params: axis
direction: [x,y,z], axis position
(in voxel grid): [x,y,z], revolute
range (degree): [min,max], slide
```

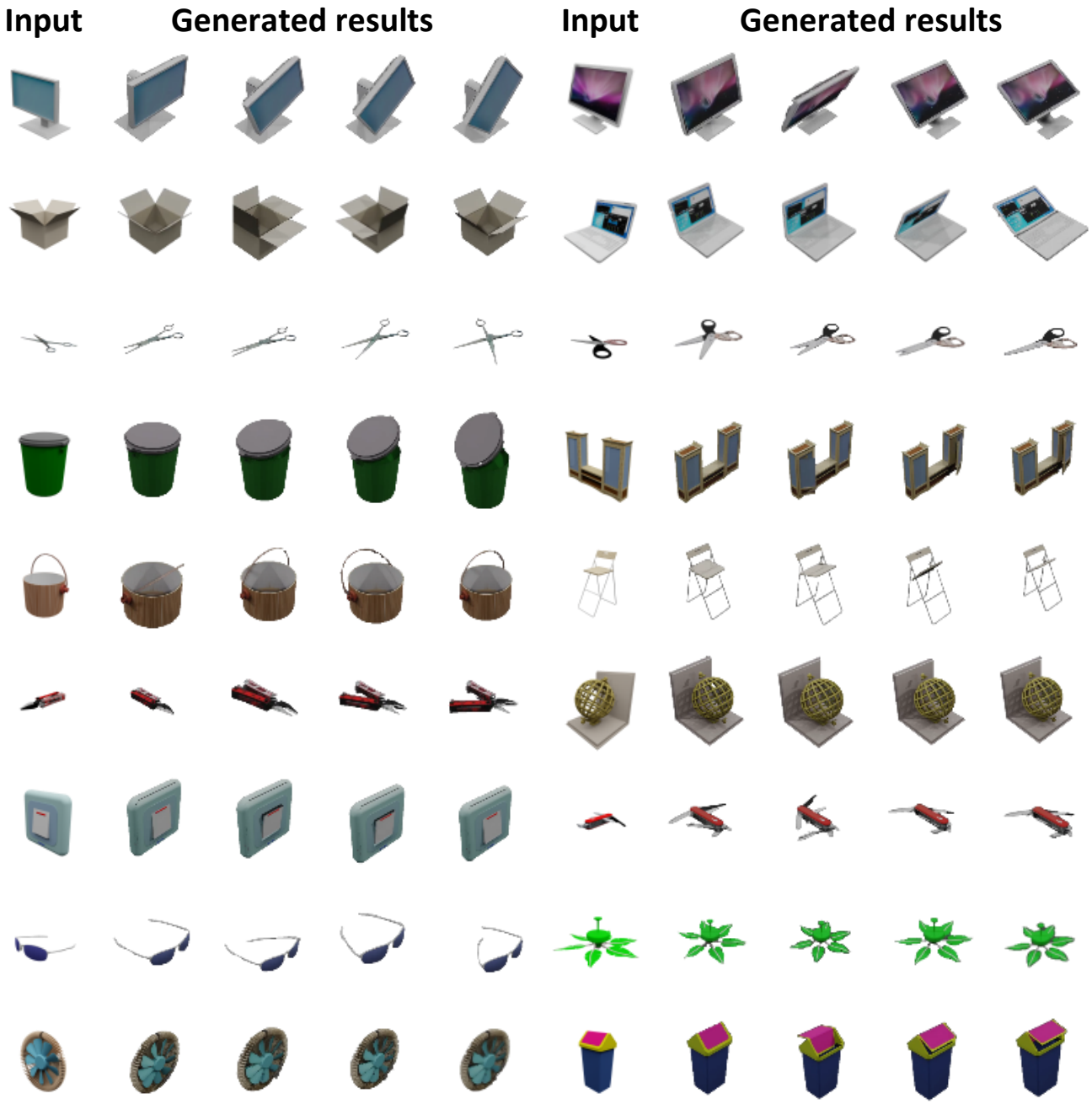

Figure 1. **More qualitative results of PhysX-Anything.** We present additional visualization results of PhysX-Anything, showing that it maintains robust performance even for complex geometries and diverse physical properties.

```
direction: [x,y,z], slide range (in
voxel grid): [min,max]
...
Note: There are Five movement
type: A (Move freely), B (Slide),
C (Rotate around axis), D (Rotate
around point), CB (Rotate and
Slide), E (Fixed to the world).
```

4.2. Prompt for geometry information of parts.

```
Based on the structured description
of l_n, generate its 3D voxel grid
in the following format (voxel
grid=32, use numbers from 0 to
```

```
32767, merge maximal consecutive
runs:  199...216 -> 199-216):  184
198 199-216 230-237...
```

5. Details of VLM-based benchmark

Since multiple factors – such as the initial pose of the object, the feasible movement range, and the underlying geometry – jointly influence the observed motion, it is inherently difficult to design a fair evaluation protocol. In particular, variations in initialization or range settings may bias the outcomes, making it challenging to construct a standardized and reliable metric that purely reflects the quality of the generated movements. Therefore, to automatically and fairly assess the performance of **kinematics** (articulation), we resort to a powerful VLM (GPT-5) to evaluate motion quality: given rendered articulation videos of the generated results, the VLM is asked to score and compare their motions.

System prompt.

```
You are given a single grid video
<GRID_VIDEO> and basic information
of the object.
Layout (2 rows x 3 columns), with
zero-based indices (row, col):
Top row:   (0,0)=GT, (0,1)=A,
(0,2)=B
Bottom row:  (1,0)=empty/ignore,
(1,1)=C, (1,2)=D

Task:
1) Slice the grid accordingly and
analyze GT: determine its category
and motion pattern.
2) For A/B/C/D, evaluate similarity
to GT in motion and geometry
(after mentally aligning via rigid
transform).
3) Ignore material/texture/color/lighting
and pure orientation/viewpoint
differences.
4) Rank A/B/C/D by similarity to
GT. Do not rely on intra-candidate
comparisons as the primary basis.
5) Return the JSON specified in the
system prompt.

Json schema:{
"A":  {"geometry_rank":  x,
"motion_rank":  x},
"B":  {"geometry_rank":  x,
"motion_rank":  x},
"C":  {"geometry_rank":  x,
"motion_rank":  x},
"D":  {"geometry_rank":  x,
"motion_rank":  x}
}
```

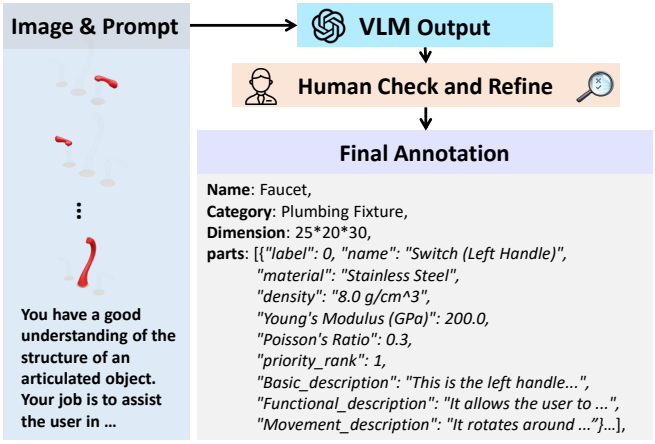

Figure 2. **Annotation pipeline for PhysX-Mobility.**

User prompt.

```
Evaluate the geometry and kinematic
reasonableness of the predicted
video against the GT video and
return exactly one JSON object per
the specified schema.
```

6. More qualitative results

Additionally, we provide more qualitative results in Fig. 1. These additional examples further demonstrate the impressive performance of PhysX-Anything in terms of both geometric fidelity and physical attributes.

7. Details of PhysX-Mobility

7.1. Annotation pipeline for PhysX-Mobility

Following PhysX-3D [2], we construct a human-in-the-loop annotation pipeline for PhysX-Mobility, as illustrated in Fig. 2. We first render part-based segmentation images and design a prompt to generate raw annotations using GPT-5. Then, human volunteer will verify and refine these raw annotations to obtain the final high-quality, fine-grained physical attribute annotations.

7.2. Statistics and Distribution of PhysX-Mobility

We report the distribution of PhysX-Mobility in fig 3, which comprises 47 common categories enriched with detailed physical annotations.

7.3. Limitation and Future Work

Our in-the-wild experiments demonstrate strong generalizability. However, limited data diversity still constrains the generation of highly detailed and diverse 3D assets, which we will address in future work.

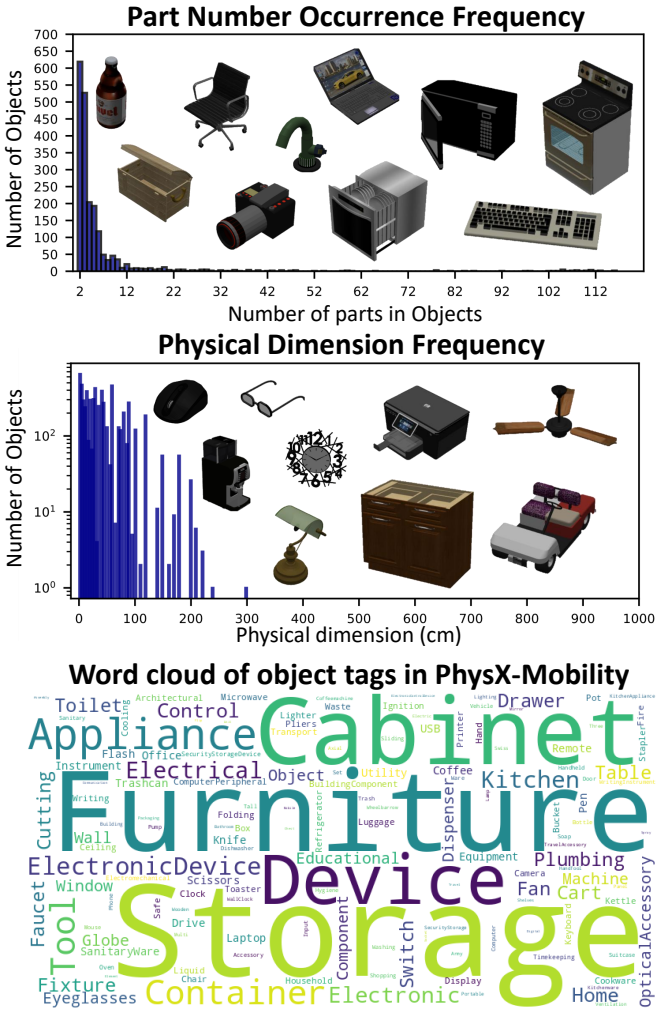

Figure 3. **Statistics and distribution of PhysX-Mobility.**

References

[1] 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. 1

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