Advancing Multimodal LLMs by Large-Scale 3D Visual Instruction Dataset Generation

Supplementary Material

Overview

Below is a summary of the contents in each section of this supplemental material:

- Sec. 1: Details of text prompt generation as image description utilized for controlled image generation.
- Sec. 2: Details of VQA prompt generation given the ground truth camera-object relations.
- Sec. 3: Details of LLM-based grading for MLLM response evaluation.
- Sec. 4: Preliminary studies on data curating for camera-object relation VQA dataset by text-image aligning metrics, and by pose estimation models.
- Sec. 5: Ablations on selections of DM backbones.
- Sec. 6: Factors influence generation quality of our image generation pipelines.
- Sec. 7: Details of user study on image quality evaluation.
- Sec. 8: More qualitative comparisons between finetuned LLaVA models to commerical SOTA models.
- Sec. 9: 48 image examples illustrate the diversity of *Ultimate3D* dataset and benchmarks on object categories, camera-object relations, and background context.
- Sec. 10: Additional discussions and insights on using synthetic dataset generated by our framework for MLLM finetuning.
- Sec. 11: Algorithm of our 3D visual instruction dataset generation pipeline.

1. Image Description Generation

We provide system prompts given to GPT-4oto generate context description for image generation. Corresponding section in main paper is Sec. 3.2. In the prompt, we provide one-shot example for better robustness of the text generation.

Specifically, generated text prompt will be merged with default positive prompt like *"detailed, 4K, 35mm photograph, professional"*, and customized camera-object prompt *"the image shows a [β] view of a [c]"*, in order to enhance image generation quality.

```
SYSTEM_PROMPT_FOR_IMAGE_GEN = "You are
    an expert in generating concise and
    diverse descriptions of an object to
    guide image generation model like
    DAll-E3. You should use your common
    sense to generate 1 sentence
    description of the given object.
    Additionally, you should also
    generate 1 sentence of a common
    real-world scene given the object.
    Both sentences of descriptions
    should not include more than one of
    the given object to avoid ambiguity.

Example Input: Please generate the
    visual prompt of chicken.
Example Output: A white and brown
    chicken standing in a cage made of
    metal bars. The chicken has a long,
    curved beak and its feathers are
    fluffy and white.
"
```

2. Text Instruction Generation

We provide the system prompt to generate QA pairs given camera-object relation as below. Corresponding section in main paper is Sec. 3.4. In the prompt, we provide few-shot example for better instructions for diversity of the generated text.

```
SYSTEM_PROMPT_FOR_VQA_GEN = "You are an
    AI visual assistant, and you are
    seeing a single image. What you see
    are provided with several sentences,
    describing the same image you are
    looking at. Answer all questions as
    you are seeing the image.

Design a conversation between you and a
    person asking about this photo. The
    answers should be in a tone that a
    visual AI assistant is seeing the
    image and answering the question.
Ask diverse questions and give
    corresponding answers.

Include questions asking about the
    visual content of the image,
    including the object orientation and
    its corresponding azimuth degree,
    the camera viewpoint of the image
    and corresponding camera elevation
    angle, the camera shot type of the
    image and the distance from camera
    to the target object. Only include
    questions that have definite answers:
(1) one can see the content and its
    orientation degree in the image that
```

the question asks about and can answer confidently;
(2) one can recognize camera viewpoint and camera shot of the image that the question asks about and can answer confidently;
The question can be multiple choice questions given the choice, or open-ended questions. For multiple choice, the option expression can be diverse but with the same meaning. For example, "front" also means "directly towards the camera"; 'back' also means 'away from the camera', etc. Use the common sense to make options diverse if possible.
Provide at most 5 pairs of question and answer pairs. Prior the confident questions. Do not ask any question that cannot be answered confidently.

You should ask questions from multiple QA templates as the below examples show.

Example Input Description:
The image shows a front view of police van. A blue and white police van with the word "POLICE" emblazoned on the side is parked on a city street. Nearby, a couple of officers are discussing a recent incident while pedestrians walk by, some glancing curiously at the vehicle.
The police van is with an azimuth of 181.518 degree, facing Front direction.
The elevation angle of the camera to the police van is 81.41 degree, the camera viewpoint is Horizontal view.
The relative distance from the camera to the police van is 1.132 meters, the camera shot type is Close-up.

Example Output Question and Answer Pairs:
Question:
From the camera's perspective, is the police van in the picture facing straight or oriented at an angle? Options: (a) Directly towards the camera (b) At an angle
===
Answer:
(a) Directly towards the camera
===
Question:
Is the police van in the picture facing the camera or away from the camera? Options: (a) Away from the camera (b) Facing the camera
===
Answer:
(b) Facing the camera
===
Question:
Which direction is the police van facing in the image? Options: (1) Back (2) Front
===
Answer:
(2) Front
===
Question:
Is the police van facing back or front from the camera's perspective? Options: (a) Back (b) Front
===
Answer:
(b) Front
===
Question:
Is the photo taken directly above the police van or from the side? Options: (a) Taken directly (b) From the side
===
Answer:
(b) From the side
===
Question:
Is the photo taken far away the police van or taken closely?
===
Answer:
The relative distance from the camera to the police van is 1.132 meters, thus it is with a close-up shot. This indicates the photo is taken closely.
===
Question:
What is the elevation viewpoint of the image? Options: (1) Top (2) Horizontal (3) Bottom
===
Answer:
(2) Horizontal
===
Question:
Which camera shots is the image? Options: (1) Close-up (2) Medium-shot (3) Long-shot
===
Answer:
(1) Close-up"

3. LLM-based Response Grading

We provide the system prompt for evaluation of the MLLM response to our camera-object relation multiple choice questions.

```
SYSTEM_PROMPT_FOR_GRADING_MLLM_RESPONSE
    ='You are a helpful and precise
    assistant for checking the quality of
    the answer. You should review all
    listed choices and compared to the
    Answer content, and judge whether the
    Answer is correct (yes), or not (no).
    You should only focus on the major
    content of the answer, not the detailed
    number or symbol. Please answer in only
    yes or no'
```

4. Metric Limitations for Evaluating Camera-Object Relation

In Fig. 1, we show an example of using ImageReward [8] for the dataset curation by evaluating the alignment between generated image and text prompts. Preliminary test indicates that text-image aligning metric is not sensitive to camera-object relation text prompts. Thus we may not rely on those metrics for data curation purpose.

Alternatively, in Fig. 2, enlighten by [3], we train a transformer-based pose estimation model on a 10-fold manner for generated images on each single category. The images with less prediction errors may indicate successful generation. However, the pose estimation model is not robust to complex objects and diverse camera-object relations in our dataset. Using pose estimation model prediction error is not robust for data curation purpose.

5. Ablation of DM Backbones

In Tab. 1, we perform quantitative comparisons on general image visual quality between different DM backbones: SD V1.5 and SDXL. We randomly sample 1,000 generated images (I_{syn}), each image corresponds to: (1) the RGB visual prior rendered using Blender (one element of I_β); (2) the text prompts ($\mathcal{T}_{img}$) provided to the diffusion model for image generation. We evaluate the generated images on the following aspects:

- Prompt following. CLIP-T (CLIP-Text Score [1]) and IR (Image Reward [8]) metrics are employed to assess the alignment between the generated images and the input prompts. The results indicate that images produced by the SDXL backbone exhibit superior prompt alignment.
- Fidelity. CLIP-I (CLIP-Image Score [1]) and DINO Score [4] are used to measure the similarity between the visual prior images and the synthesized outputs. Findings demonstrate that SDXL more effectively preserves fidelity to the visual prior.
- Realism. FID [2] is used to compare the distributions of generated images with a subset of LAION-aesthetic [5] images, revealing that SDXL significantly surpasses SD V1.5 in terms of image quality and realism.

Ablations	CLIP-T ↑	CLIP-I ↑	DINO ↑	IR ↑	FID ↓
Ours (SD-V1.5)	29.80	83.56	69.36	-0.29	289.99
Ours (SDXL)	**34.40**	**87.44**	**70.79**	**0.23**	**256.94**

Table 1. **Ablations on DM Backbones**. SDXL provides better fidelity (CLIP-I, DINO), realism (FID), and text-image alignment (CLIP-T, IR) than other alternatives.

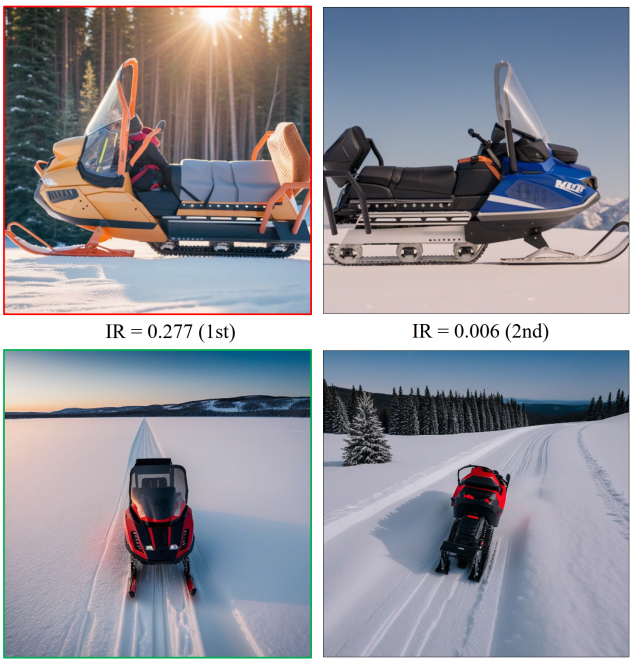

Figure 1. **Limitations of Text-Image Aligning Metrics.** We evaluate the ImageReward [8] (IR) between the text promopt regarding object orientation with above 4 synthetic images. The image matches the correct camera-object relation text prompt ("front") may reach the highest score. However, image with the highest matching score (it faces to "left" direction) is not the correct one with expected orientation. This failure is prevailing across all categories.

6. Factors influencing Image Generation

Several factors may lead to unreasonable artifacts in our image generation pipeline. In Fig. 3, the upper-left example shows that bottom-view camera-object relations can introduce physical anomalies, such as the tricycle appearing with its front lifted unnaturally. The upper-right example high-

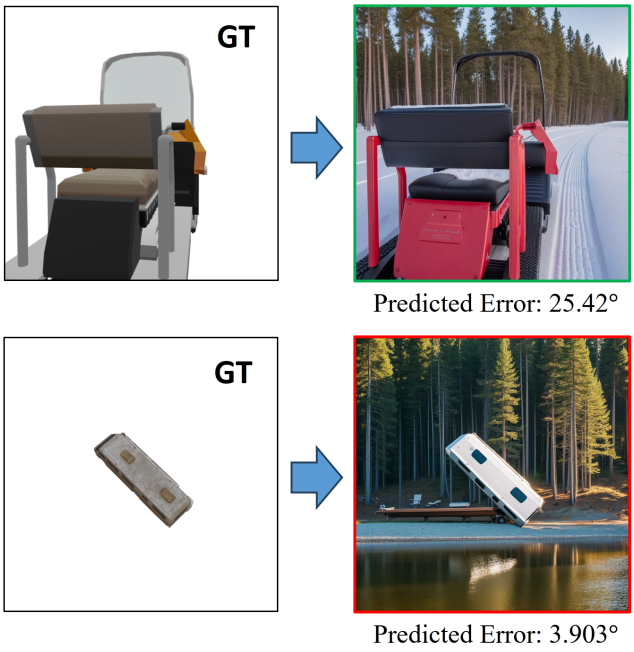

Figure 2. **Failures of Pose Estimation Model.** Enlighten by [3], we train a transformer-based pose estimation model on a 10-fold manner for generated images on each single category. The images with less prediction errors may indicate successful generation. However, the pose estimation model is not robust to complex objects and diverse camera-object relations in our dataset.

lights that low-quality 3D assets can produce fragmented depth maps or Canny edge visual priors, resulting in unrealistic generated images. The lower-left example shows that small subjects viewed from a long-shot perspective may lead to duplicate subjects in the generated image. The lower-right example illustrates how complex 3D asset structures can cause texture blending issues, particularly where intricate geometry is present.

There are several ways to reduce the influences from those factors. First, choosing 3D assets with reasonable number of meshes, and with plausible textures may filter the low-quality and texture blending failures. Second, using simple and compact image description prompts to generate long-shot camera-shot images will significantly reduce the duplicate subject failures. Reduce the image generation with high elevation angle from bottom view will significantly reduce physical anomaly.

7. User Study Details

The Fig. 4 shows the UI page of our user study. We randomly selected 200 images from the *Ultimate3D* dataset. Each image corresponds to an RGB prior generated by Blender. During the user study, users will see a pair of side-by-side images. One is an RGB image generated by Blender, and the other is the synthetic image generated by DMs using that RGB image as guidance prior. A vivid preserving of 3D geometry and structure in the synthetic image is a success. Users are required to provide a binary answer regarding the success of the image generation. Each image pair will be manually reviewed by at least 3 users.

8. Additional Qualitative Results

Based on *Ultimate3D* benchmark, we provide additional Fig. 5 to show more qualitative comparisons between finetuned LLaVA model to commercial SOTAs. The finetuned LLaVA model outperforms other models.

9. Additional Visuals of *Ultimate3D* Dataset

In Fig. 6, we shows more examples of diversity on object categories, camera-object relation, and background contexts.

10. Additional Discussions

We provide some insights of using synthetic generated visual instruction dataset for MLLM finetuning. For the task of camera-object relation recognition for 100 categories (as we collected in *Ultimate3D* dataset), a dataset comprising 100K to 1M VQAs will significantly improve baseline MLLMs by one-epoch finetuning. Further scaling up the dataset volume, finetuneing epochs, or model parameters may yield saturation.

Moreover, the quality of synthetic dataset is crucial to the improvement margin. Using dataset with higher success rate (ours is 93.07% as main paper Sec. 4.5) provides better performances than using dataset generated other alternative backbones (i.e. SD-V1.5).

The coverage of arbitrary camera-object relation is also important for generalization. Our preliminary test using real images which are highly biased to front-facing direction from Pascal3D+ [7], results in poor generalization on uncovered orientations and viewpoints. This illustrates the importance of using synthetic dataset to eliminate the dataset biases.

For numerical prediction of camera-object relation, our preliminary test shows plausible success using MEBOW [6] dataset to improve MLLMs on predicting human orientation angles. The average degree prediction accuracy is competitive to the original paper [6]. However, the extended numerical prediction of 100 categories and 3 types of camera-object relations may require even larger dataset volume than *Ultimate3D*. Our framework is able to scale up the dataset volume. But due to the limitation of computing power, we leave it as interesting future work.

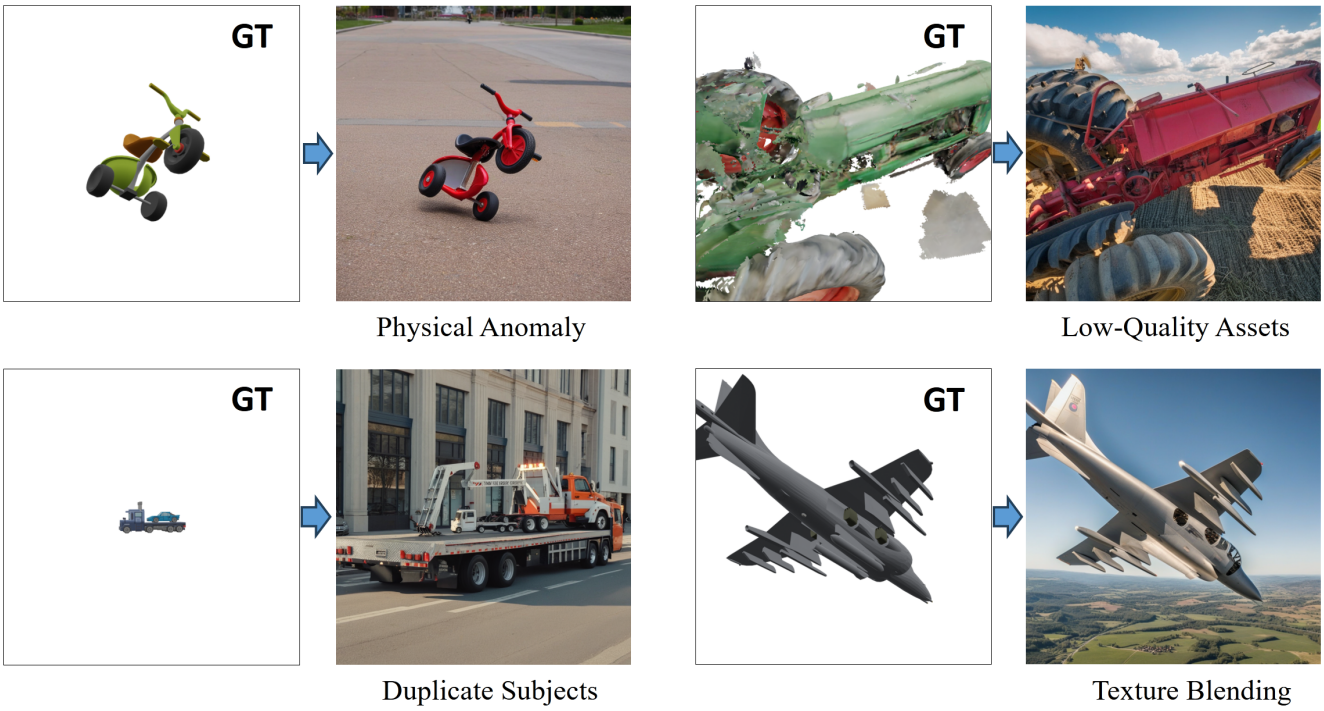

Figure 3. **Factors influencing Image Generation.** The upper-left example shows that bottom-view camera-object relations can introduce physical anomalies, such as the tricycle appearing with its front lifted unnaturally. The upper-right example highlights that low-quality 3D assets can produce fragmented depth maps or Canny edge visual priors, resulting in unrealistic generated images. The lower-left example shows that small subjects viewed from a long-shot perspective may lead to duplicate subjects in the generated image. The lower-right example illustrates how complex 3D asset structures can cause texture blending issues, particularly where intricate geometry is present.

11. Algorithm of 3D visual instruction dataset generation pipeline

Algorithm 1 Synthetic VQA generation

Input: 3D asset A, asset category c, camera-object relation β.

Parameter: Renderer $\mathcal{R}$, DM-based image generator $\mathcal{G}$, image decoder $\mathcal{D}$, LLM text generator $\mathcal{L}$, system prompt given to $\mathcal{L}$ for generating image context description p_{img}, and for generating QA pairs p_{qa}, DM sampling steps T.

Output: Generated synthetic image I_{syn}, corresponding text QA pairs $\mathcal{T}_{qa}$.

$I_\beta = \mathcal{R}(A, \beta)$ ..... # Render 3D priors
$\mathcal{T}_{img} = \mathcal{L}(c, p_{img})$ .... # Generate image description
$z_T \sim \mathcal{N}(0, I)$
for $t = T, \cdots, 1$ **do**
    $z_{t-1} = \mathcal{G}(z_t, I_\beta, \mathcal{T}_{img})$ ..... # DM denoising
end for
$I_{syn} = \mathcal{D}(z_0)$
$\mathcal{T}_{qa} = \mathcal{L}(c, \beta, p_{qa})$ ..... # Generate QA pairs
return $I_{syn}, \mathcal{T}_{qa}$

References

[1] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 3

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[3] Wufei Ma, Qihao Liu, Jiahao Wang, Angtian Wang, Xiaoding Yuan, Yi Zhang, Zihao Xiao, Guofeng Zhang, Beijia Lu, Ruxiao Duan, et al. Generating images with 3d annotations using diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 4

[4] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[5] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information*

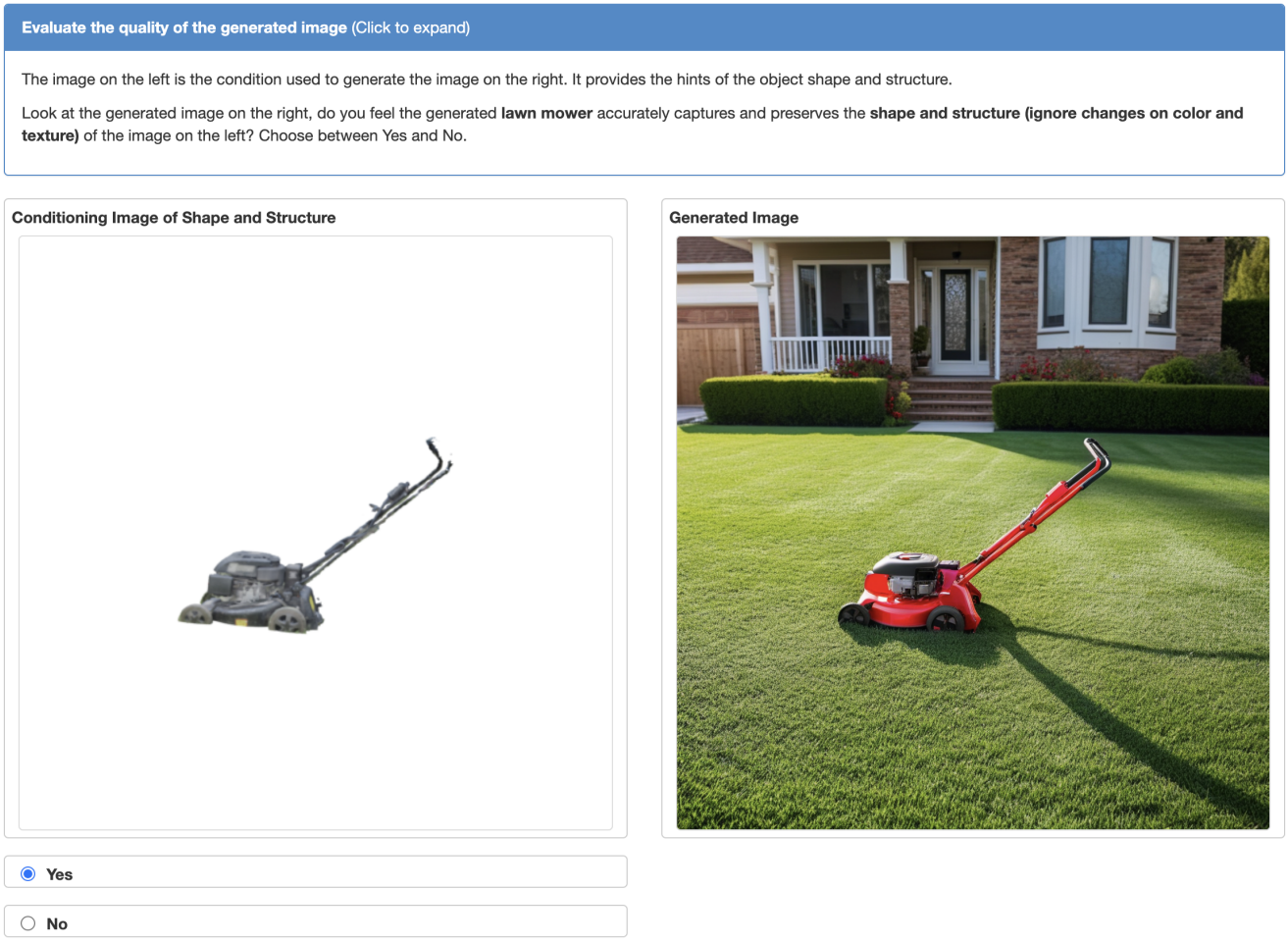

Figure 4. **User Study UI Page.** Users will see a pair of side-by-side images. The left hand side is an RGB image generated by Blender, and the right hand side is the synthetic image generated by DMs using that RGB image as guidance prior. A vivid preserving of 3D geometry and structure in the synthetic image is a success. Users are required to provide a binary answer regarding the success of the image generation.

Processing Systems, 35:25278–25294, 2022. 3

[6] Chenyan Wu, Yukun Chen, Jiajia Luo, Che-Chun Su, Anuja Dawane, Bikramjot Hanzra, Zhuo Deng, Bilan Liu, James Z Wang, and Cheng-hao Kuo. Mebow: Monocular estimation of body orientation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3451–3461, 2020. 4

[7] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014. 4

[8] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

Figure 5. **Additional Qualitative Results.** We show the evaluations of camera-object relation recognition capability of GPT-4o, Claude-3-Sonnet, LLaVA-1.6-13B, and finetuned LLaVA-1.6-13B, on *Ultimate3D* benchmark. Each model is asked the questions (in gray boxes) regarding object orientation, camera viewpoint, and camera-shots type, together with the input images on the left. The model responses illustrate that finetuned LLaVA-1.6-13B model by *Ultimate3D* dataset significantly outperforms all compared models.

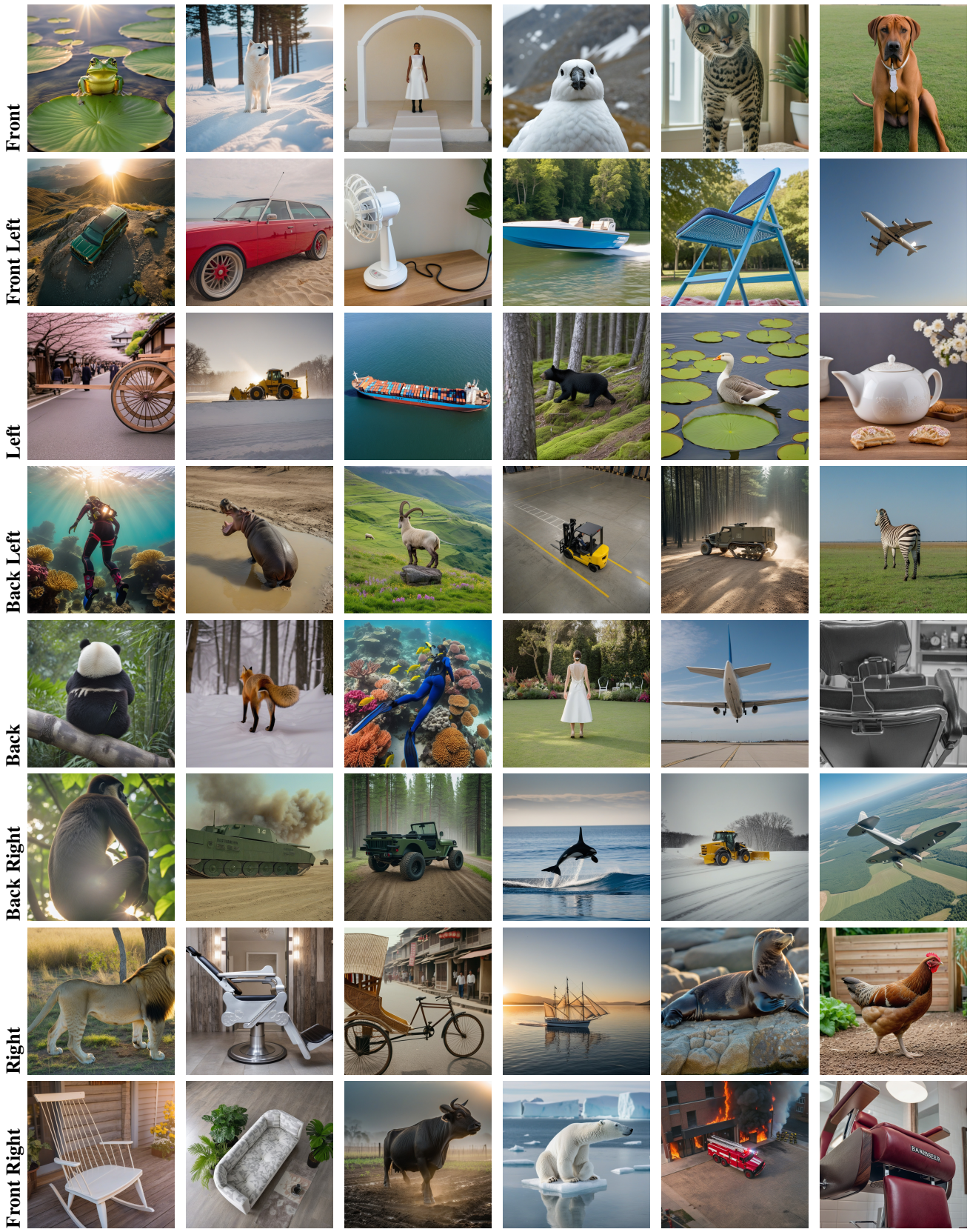

Figure 6. **Diversity of *Ultimate3D***. Our *Ultimate3D* dataset and benchmark cover 100 categories of objects, range diverse camera-object relation settings, and provide plausible image quality. (Each row shows images with the same orientation but in diverse subject and context.)