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

Liu He[1] **Xiao Zeng**[2] **Yizhi Song**[1] **Albert Y. C. Chen**[2] **Lu Xia**[2] **Shashwat Verma**[2]
Sankalp Dayal[2] **Min Sun**[2] **Cheng-Hao Kuo**[2] **Daniel Aliaga**[1]
[1] Purdue University [2] Amazon
{he425, song630, aliaga}@cs.purdue.edu
{zenxiao, aycchen, luxial, shashwv, sankalpd, minnsun, chkuo}@amazon.com

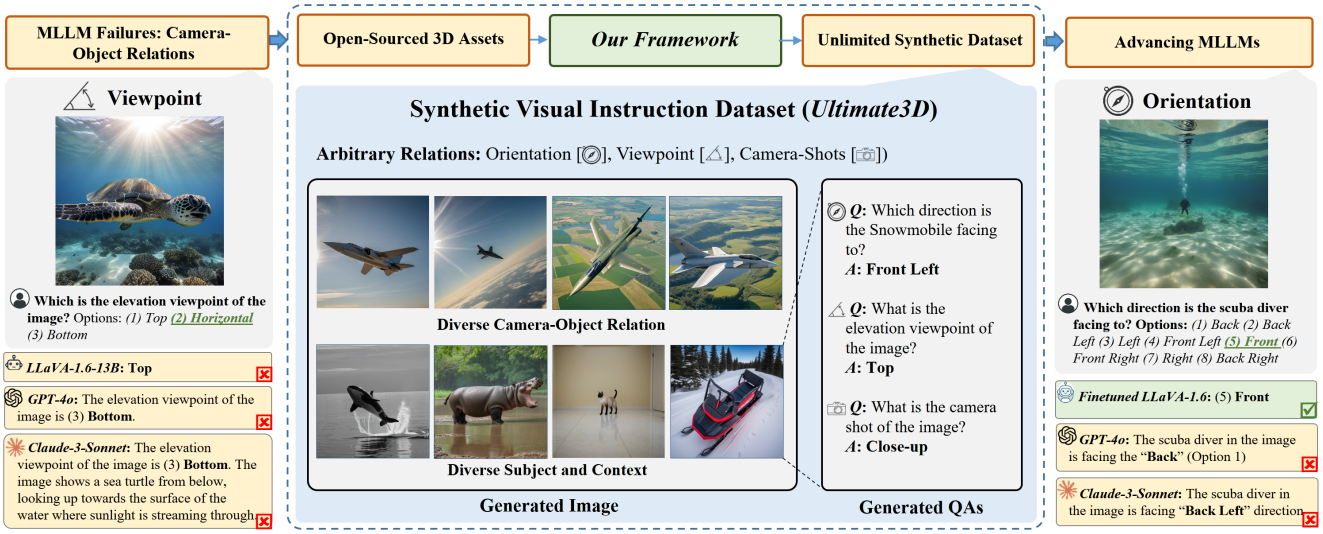

Figure 1. **Generating Synthetic Visual Instruction Dataset**. Our framework uses open-sourced 3D assets to generate photo-realistic images with precisely controlled camera-object relation. Corresponding text instructions are also generated by Large Language Model (LLM). The generated *Ultimate3D* dataset (240K) and benchmark (8K) advance baseline MLLM models (LLaVA-1.6, Llama-3.2-Vision, etc.) to outperform commercial MLLMs (GPT-4o, Claude-3-Sonnet, etc.) on camera-object relation recognition tasks.

Abstract

*Multimodal Large Language Models (MLLMs) struggle with accurately capturing camera-object relations, especially for object orientation, camera viewpoint, and camera shots. This stems from the fact that existing MLLMs are trained on images with limited diverse camera-object relations and corresponding textual descriptions. To address this, we propose a synthetic generation pipeline to create large-scale 3D visual instruction datasets. Our framework takes 3D assets as input and uses rendering and diffusion-based image generation models to create photorealistic images preserving precise camera-object relations. Additionally, large language models (LLMs) are used to generate text prompts for guiding visual instruction tuning and controlling image generation. We create Ultimate3D, a dataset of 240K VQAs with precise camera-object annotations, and corresponding benchmark. MLLMs fine-tuned on our proposed dataset outperform commercial models by a large margin, achieving an average accuracy improvement of **33.4%** on camera-object relation recognition tasks. Our code, dataset, and benchmark will contribute to broad MLLM applications.*

1. Introduction

The advent of Multimodal Large Language Models (MLLMs) [5, 8, 32, 33, 37, 40, 41, 47, 64] have demonstrated remarkable success on tasks including image captioning, embodied robotic manipulation, and visual question answering (VQA). MLLMs bridge visual understanding from vision encoders [11, 50] with powerful verbal capability from large language models (LLMs) [7, 66]. Collectively this provides end-to-end inferencing across image and text modalities. In fact, the recently released GPT-4o [49] and Claude-3.5-Sonnet [1] have extended their performances to be comparable to human intelligence in some

of the experimented tasks [71].

However, research works have found needs and weaknesses of MLLMs [14, 65, 76]. Tong *et al*. [65] point out the deficiency in visual representations learnt by ViT [11]. In particular, two weakness categories (i.e., orientation/direction, viewpoint/perspective) focus on camera-object relations such as distinguishing simple concepts about object orientation (”right”, ”front”, ”back”, etc.), camera viewpoint location (”top”, ”bottom”, etc.), and camera-shot styles (”close-up”, ”long-shot”, etc.). As two examples in Fig. 1 show, current SOTA model (i.e., GPT-4o) struggle with understanding such camera-object relations. The challenges to improving 3D-aware MLLM abilities on recognizing camera-object relations include: (1) capturing the complexity of the task, and (2) having a dataset containing millions of image-text pairs as is a common requirement for training MLLM on spatial perception tasks [3, 6, 17].

Nevertheless, to the best of our knowledge, none of the existing MLLMs or training datasets focus on the aforementioned camera-object relations. Prior methods for MLLM training data generation can be divided into the following two groups: (1) **Relabeling of existing real images**. These works focus on generating textual captions by deep learning models given existing images [4, 12, 34, 55, 79, 80]. Specifically, SpatialRGPT [6] and SpatialVLM [3] address spatial perception and reasoning weakness by providing a comprehensive data annotation pipeline leveraging multiple image segmentation, depth estimation, and camera calibration models to produce dense predictions for real images. But, in practice the accuracy of the predicted labels is not high and the dataset size and distribution is constrained by existing image datasets which predominantly only contain ”front” view objects. (2) **Generating synthetic images**. Typically text-to-image DMs are utilized for synthetic image generation based on provided text prompts [13, 35, 43]. But, their score-based data curation is not robust to camera-object relations, or generation failures (See Supp Sec. 4). Alternatively, simulation-based generation pipelines for visual reasoning [16, 29, 42] may provide ground truth camera-object relations, but the generated image quality, complexity, and diversity are still far from realistic. In summary, existing synthetic dataset generation pipelines fail to provide image-text pairs with ground truth camera-object relation, realistic image quality, and diverse image subject and context.

In this work, we propose a synthetic generation pipeline to create unlimited image-text pairs focusing on camera-object relations. With 3D assets (e.g., 3D models with texture, materials, and animation) as input, our framework utilizes a renderer (e.g., Blender) with arbitrary camera-object relation parameters to render a series of 3D ground truth priors (i.e., RGB, depth, segmentation). Those 3D priors are used as comprehensive conditions for stacked diffusion-based generation models (e.g., ControlNet) to generate photorealistic images preserving precise camera-object relations. Simultaneously, LLMs are used to generate versatile text prompts for the controlling of image context and object appearance, and to generate diverse VQA prompts regarding camera-object relations. With our framework, we have created the *Ultimate3D* dataset as a future benchmark for camera-object relations. MLLMs fine-tuned by our dataset explicitly outperform commercial models by a large margin (i.e., an average 33.4% accuracy improvement) on recognizing camera-object relations.

Our contributions include:

- A synthetic generation pipeline for unlimited image-text pair creation of diverse camera-object relations.
- A dataset and benchmark for further improving and evaluating camera-object relations in MLLMs.
- Improved open-sourced MLLMs (e.g., LLaVA) that outperform commercial SOTA (e.g., GPT-4o) in camera-object relation tasks.

2. Related Works

2.1. Datasets and Benchmarks for MLLMs

MLLMs use various dataset recipes for visual instruction tuning [5, 8, 32, 33, 37, 40, 41, 64]. Typically, dozens of datasets, focusing on different areas, are combined to enable the comprehensive abilities of MLLMs. Datasets include image captioning [38, 72, 73], visual reasoning [28, 39], general VQA data [19, 41, 67] on knowledge grounding [44, 46], image understanding [15], and OCR [48, 56]. Further, most visual instruction datasets are relabeled or created from other existing image-based datasets (e.g., COCO [38], LAION [54], LVIS [18], Visual Genome [30], ADE20K [81]). As far as we know, none of the datasets focus on camera-object relations. Thus, the missing ground truth camera-object relation data exacerbates the task for deep learning models, such as SpatialRGPT [6] and SpatialVLM [3]. In fact, multiple papers have indicated MLLMs' poor distinguishing ability on camera-object relations [14, 65]. Our paper fills this gap by providing a generation pipeline, dataset, and benchmark.

In particular for the visual instruction data generation, LLMs, like GPT-4o [49], have been successfully utilized in LLaVA [39, 41] for dialogue text instruction generation, and used in MMVP [65] for answer grading. However, generating high-quality images corresponding to proposed text prompts has not been well developed. In this project, we leverage the diverse text generation capability of GPT-4o for the text prompt generation and we provide an image generation pipeline regarding 3D-awareness.

2.2. Simulation-based Image Generation

CLEVR [29] and Kubric [16] use 3D engines to generate synthetic images by rendering 3D primitives at multiple

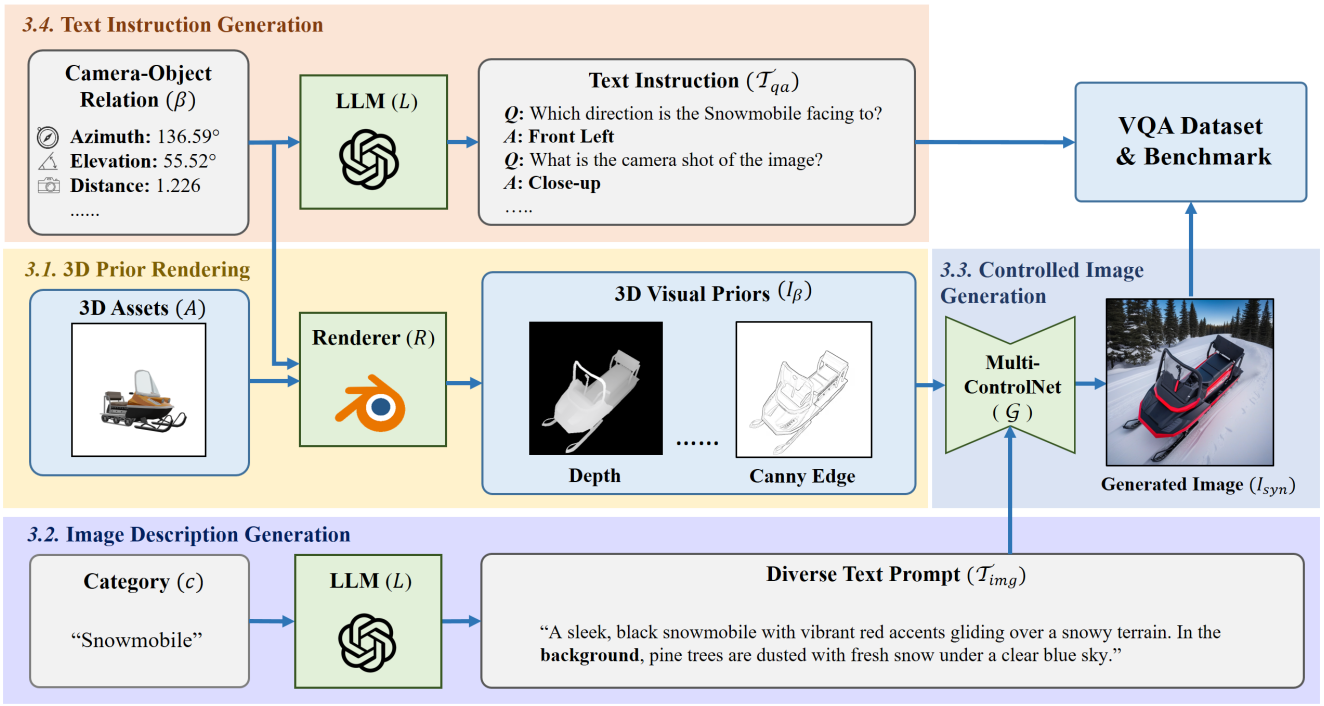

Figure 2. **Our Framework.** Each shade box corresponds to a section in Sec. 3. Given open-sourced 3D assets, our approach leverages a 3D Renderer to generate 3D visual priors (I_β) preserving ground truth camera-object relation (β). Meanwhile, LLMs take 3D asset category to generate diverse image descriptions ($\mathcal{T}_{img}$) as conditional guidance. Both the 3D visual priors and diverse text prompts are used to generate synthetic images (I_{syn}) by multiple ControlNet-based networks. Corresponding text QA instructions ($\mathcal{T}_{qa}$) are generated by LLMs given the ground truth camera-object relation. Our *Ultimate3D* dataset and benchmark (i.e. pairs of $\mathcal{T}_{qa}$ and I_{syn}) contribute to fine-tuning and evaluation of MLLMs.

arbitrary camera-object relations. Generally, simulation-based image rendering easily preserve ground truth camera-object relations in arbitrary viewpoint settings. However, the generated images are neither photorealistic nor diverse. Realism is limited by the assets' level of detail and textures, and diversity is restricted by the small number of assets. These limitations constrain the usage of synthetic images in MLLM training to only object counting or simple spatial reasoning tasks [64]. In this project, we leverage the 3D geometry prior rendered by a 3D renderer to precisely guide the image generation pipeline for preserving ground truth camera-object relation.

2.3. Generative Images

Generative models provide advanced image generation by iteratively restoring the original data by denoising from Gaussian noise [20–24, 26, 27, 36, 57–63, 74, 75]. Open-sourced Stable Diffusion [52], and SDXL [51] enable high-resolution and photo-realistic image generation and editing. While generative models provide versatile context and subject, representative works (e.g., AURORA [31], EditBench [68], Prompt-to-Prompt [25], and MagicBrush [77]) fail to control camera-object relations in the generated content, and quality guarantees rely on human checking. ControlNet [78], built on versatile backbone DMs (e.g., SDXL), provides spatial control in the generated images by using various image priors (e.g., Canny edges, semantic segmentation, relative depth).

In particular, Ma et al. [45] explore using synthetic image generation for 3D pose estimation by conditioning diffusion models on pose angles. However, their approach does not address advancing MLLMs 3D spatial understanding, nor does it consider VQA instruction generation. Moreover, their reliance on 2D Canny priors (i.e., silhouette) brings severe artifacts for images with backward viewpoint and complex depth variations. In contrast, our method incorporates depth priors to enhance geometric accuracy and leverages SDXL's superior fidelity to improve image realism. Further discussions and user studies are in Sec. 4.5 and Sec. 4.6.

3. Method

Our framework is described in Fig. 2 and Supp Sec.11. Given input 3D asset A, its category c, and arbitrary camera-object relation β, our framework generates photorealistic image I_{syn} preserving β, and corresponding text QA instruction $\mathcal{T}_{qa}$. Our pipeline consists of 3D visual prior rendering (Sec. 3.1), diverse image description generation (Sec. 3.2), controlled image generation (Sec. 3.3), and 3D-aware text instruction generation (Sec. 3.4).

3.1. 3D Visual Prior Rendering

Our system renders several 3D visual priors including 2D depth images, Canny-edge images, and segmentation masks from arbitrary viewpoints. These 3D visual priors con-

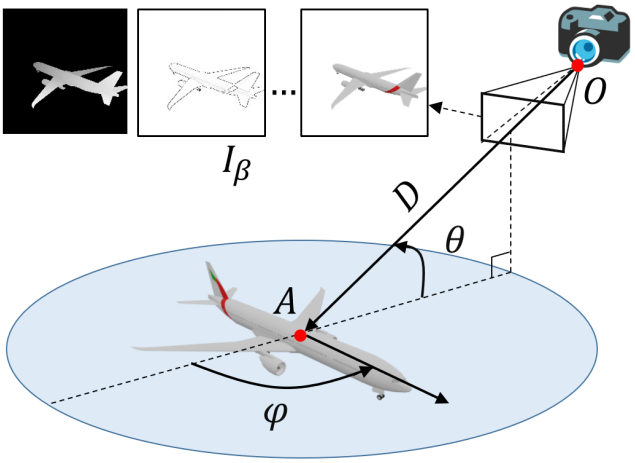

Figure 3. **3D Visual Prior Rendering.** With a general camera model, our method utilizes a 3D Renderer (e.g., Blender) to render multiple 3D visual priors given arbitrary camera-object relations.

tain the scene's geometry and will be used together with LLM-generated image descriptions to collectively control the DM-based image generation in Sec. 3.3.

Camera-Object Relation Definition. As Fig. 3 shows, we propose a simple camera model to define camera-object relations – a component which is missing in [45]. Given a 3D asset represented by the centroid A and a camera represented by aperture center O, we assume the camera is directly facing the asset. Any location/direction relation between camera and captured object can be represented by $\beta = \{\varphi, \theta, D\}$ consisting of three components:

- The **object orientation angle** $\varphi \in [0, 2\pi]$, which is the counterclockwise azimuth of the 3D asset's facing direction with respect to the camera-facing direction $\overrightarrow{OA}$.
- The **viewpoint angle** $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, which is the elevation angle of the camera over the horizontal plane of the asset.
- The **camera-object distance** $D \in [0, \infty)$, which is the length of the line $\overline{OA}$.

Visual Prior Rendering. Given an arbitrary β, the system will render a corresponding set of 2D images I_β (Fig. 3) of the 3D asset. The system renders I_β using 3D renderer R. R uses as the default camera intrinsic a perspective camera with focal length 35mm, and a default extrinsic provided by β. We keep the 3D asset in upright position and located at the center of the camera frame. Images captured at arbitrary β will result in various viewpoints, object orientations, and camera-shot distances. Specifically, the renderer R generates images of RGB data, semantic masks, and pixel-level depth maps. The rendered RGB images are further processed to provide Canny edges as additional visual priors.

3.2. Diverse Image Description Generation

Our approach uses GPT-4o as a LLM to generate diverse text prompts for image descriptions ($\mathcal{T}_{img}$) for DM-based image generation; an example is in Fig. 2. Our framework uses system prompting and few-shot prompting tricks which are broadly utilized for image captioning and QA composing [39, 41].

For the generation of $\mathcal{T}_{img}$, the LLM is guided to generate diverse descriptions of background context and object details given the object category c. The diversity and rationality of the description directly influences the image generation quality (see Supp Sec. 9). More details are in Supplemental Sec. 1.

3.3. Controlled Image Generation

We utilize DMs to generate photorealistic images I_{syn} given comprehensive visual prior I_β. A successful generation of I_{syn} shall preserve ground truth camera-object relation information, and be of high fidelity. We first describe a traditional text-to-image DM pipeline, extend it to stacked multiple ControlNets using various images of I_β, and obtain precisely-controlled image generation.

Text-to-Image Generation. We use text-to-image DM (i.e., SDXL-1.0 [51]) as our backbone. The denoising U-Net, $\mathcal{G}_\theta$ parameterized by θ, conducts iterative denoising diffusion steps of Gaussian noise ϵ for a total of T time steps. A text-encoder transforms the text prompt $\mathcal{T}_{img}$ into text embedding for guidance injection. The iterative denoising is defined as:

$$z_{t-1} = \mathcal{G}_\theta(\mathcal{T}_{img}, z_t, t), t = T, ..., 1 \quad (1)$$

where $z_T = \epsilon \sim \mathcal{N}(0, I)$, and z_0 is the latent variable of the generated image. It can be decoded to an RGB image by pretrained image decoder $\mathcal{D}$ as $I_0 = \mathcal{D}(z_0)$.

Precise Controlling via Multiple ControlNets. ControlNet [78] guides image generation by injecting additional controlling features derived from various 2D image visual priors. There are dozens of off-the-shelf ControlNets, each pretrained by a single type of visual prior. In our work, we leverage multiple ControlNets to capture comprehensive visual priors from I_β (depth, Canny edge, etc.) where each one is weighted by coefficients w. The denoising process using multiple ControlNets ($\mathcal{C}$) is defined as:

$$z_{t-1} = \mathcal{G}_\theta(\mathcal{T}_{img}, z_t, t, \sum_{k=1}^{|\beta|} w_k \cdot \mathcal{C}^k(I_\beta^k)), t = T, ..., 1. \quad (2)$$

Similarly, the output generated image $I_{syn} = \mathcal{D}(z_0)$ and I_{syn} is expected to have ground truth camera-object relations.

3.4. Text Instruction Generation

We utilize GPT-4o as a LLM to generate QA pairs on camera-object relation ($\mathcal{T}_{qa}$) for visual instruction tuning (more details in Supplemental Sec.2).

Specifically, the LLM is prompted to generate QA pairs given the camera-object relation β. We subdivide each component of β into categorical types for effective referring through textual response by MLLMs:

- Object orientation angle φ is classified into eight object orientation types, each covers a bin of $\frac{\pi}{4}$ azimuth (i.e. "right", "front right", "front", "front left", "left", "back left", "back", "back right").
- The viewpoint angle θ is classified into three viewpoint types, each covering a range of $\frac{\pi}{3}$ elevation angle (i.e. "horizontal", "top", "bottom").
- The camera-object distance D is in relative units of Blender. It reflects the ratio of object dimension over the full picture size of the camera ($D = 1$ indicates fully covered). The three camera-shot types (i.e. "close-up", "medium-shot", "long-shot") is decided by two cutoff values 1.25 and 3.0.

Examples of generated multiple-choice style QA pairs as MMVP benchmark [65] are showed in Fig. 2. For question generation, we prompt LLMs to generate template questions for each prediction task of object orientation, viewpoint type, and camera-shot type. We list all possible options, including the ground truth β, in random order. MLLMs are expected to choose the option according to β.

4. Experiments

4.1. Visual Instruction Generation

Data Source. Our framework uses open-sourced 3D assets datasets from ObjaverseXL [9, 10] and ShapeNet [2]. Partial 3D assets are labeled following the synset list from ImageNet-1K [53]. We leverage this subset and define the asset category by its synset label. In particular, we manually select 100 synsets with explicit "facing" orientation from the original list of ImageNet. In total, there are 1196 assets used for synthetic visual instruction generation.

ControlNets. We utilize SDXL-1.0 [51] as the backbone of ControlNets. For selection of multiple visual priors, we found using both Canny edges and relative depth provides the best mix of camera-object relation preservation and image quality. The diffusion step $T = 30$, and relative weights for depth and Canny edge ControlNets are 0.5 and 0.8, respectively. Several seconds are needed to generate an image by a single H100 GPU. Ablations and user studies are shown in Sec. 4.5 and Sec. 4.6.

4.2. Ultimate3D Dataset and Benchmark.

Our framework creates the *Ultimate3D* dataset consisting of 85K synthetic images and 18K cropped images of a single human per photograph and corresponding human orientation labels from MEBOW [69]. For the synthetic images, we use 1180 3D assets covering 100 categories. For each 3D asset, there are 72 possible camera-object relations (8 orientations, 3 viewpoints, and 3 camera shot types) and we render a corresponding synthetic image for each relation. Then, each synthetic image is described by up to 3 textual QAs corresponding to object orientation, camera viewpoint, and camera-shot type, yielding a total of 240K corresponding QAs. Examples of synthetic images are in Supp Sec. 9. The images from MEBOW are useful because otherwise there is a strong lack of 3D assets of human which we find crucial to fine-tuning. The MEBOW images only have one QA corresponding to human orientation.

The *Ultimate3D* benchmark (Fig. 4) has two sets: (1) Synthetic Set. It contains 1200 synthetic images and 3600 QAs generated by our framework. Each image and QA has been manually reviewed to ensure correctness. This set covers all 100 categories with explicit facing orientation. The 3D assets utilized for benchmark generation do not overlap with those used for dataset generation. (2) Real Set. It contains two portions. The first portion is 2443 real images collected from Pascal3D+ [70] covering 10 categories (e.g. aeroplane, bicycle, etc.). Each image corresponds to a QA based on provided labels of object orientation or camera viewpoint. The second portion is 800 real images of humans collected from the MEBOW dataset [69]. Each image corresponds to a QA on human orientation based on provided labels. In total, there are 3243 VQAs in the Real set. We resample the real images to keep a uniform distribution for each orientation and viewpoint type. Since no camera-shot information is provided in the real image set, the QAs in the real set exclude camera-shot type questions.

Both the *Ultimate3D* dataset and the benchmark have no camera-object relation bias, which is very uncommon in the real image datasets (e.g. most images capture objects facing straight to the camera [70]). In contrast, in our dataset the number of images in each orientation, viewpoint, and camera-shot type are **evenly** distributed.

4.3. Advancing 3D-Aware MLLMs

We explore how much our synthetic *Ultimate3D* dataset can improve the MLLMs recognition capability on camera-object relation. The evaluation results on both synthetic and real image VQA benchmarks indicate the fine-tuned baseline models are significantly improved from random guess and outperform commercial SOTAs.

Training Details. We choose LLaVA-1.5/1.6 [40] with both 7B and 13B parameters, and Llama-3.2-Vision-11B [47] as baselines. Each model is fine-tuned on the *Ultimate3D* dataset for one epoch. Specifically, both the MLP of vision-language connector and the LLM decoder are trained, while the vision encoder is frozen. All fine-tuning hyperparameters follow the official instructions from [40, 47]; training is 12 hours on 4 H100 GPUs.

Evaluation Settings. We evaluate the baseline MLLMs and their fine-tuned versions on *Ultimate3D* benchmark and

Table 1. **Quantitative Comparisons.** We fine-tune LLaVA models by *Ultimate3D* dataset, then evaluate the MLLM response accuracy (%) on *Ultimate3D* and *MMVP* benchmarks (*MMVP* is a public benchmark showing our cross-dataset capability). Fine-tuned LLaVA models outperform SOTAs by an average of 33.4% among all three tasks.

MLLMs	**Orientation** (% ↑)				**Viewpoint** (% ↑)				**Camera-Shots** (% ↑)	
	Ultimate3D			*MMVP*	*Ultimate3D*			*MMVP*	*Ultimate3D*	*MMVP*
	Both	Syn	Real		Both	Syn	Real		Syn	
LLaVA-1.5-7B	18.3	11.1	22.5	25.0	31.8	38.4	25.3	29.2	46.0	66.7
LLaVA-1.6-7B	18.2	14.7	20.3	25.0	36.8	43.9	29.8	33.3	46.6	71.4
LLaVA-1.5-13B	17.9	18.7	17.4	22.7	38.9	35.0	42.8	41.7	44.6	61.0
LLaVA-1.6-13B	16.1	16.4	16.0	15.9	30.7	30.3	31.1	33.3	42.3	61.9
Llama-3.2-V-11B	5.6	5.3	5.7	6.8	19.8	26.2	13.6	38.1	29.6	25.0
Finetuned LLaVA-1.5-7B	68.8	85.6	58.8	54.5	70.6	81.3	60.0	66.7	94.0	66.7
Finetuned LLaVA-1.6-7B	71.5	86.9	62.4	61.4	71.3	**83.3**	59.5	66.7	94.1	50.0
Finetuned LLaVA-1.5-13B	70.0	85.6	60.8	50.0	69.7	81.9	57.7	66.7	94.2	**76.2**
Finetuned LLaVA-1.6-13B	72.4	**88.1**	63.1	**65.9**	**72.3**	83.0	**61.8**	**75.0**	**94.8**	71.4
Finetuned Llama-3.2-V-11B	**74.2**	87.3	**66.4**	50.0	69.1	79.6	58.7	58.3	93.1	**76.2**
GPT-4o	43.1	51.0	38.5	40.9	54.1	63.5	44.9	70.8	41.7	42.9
GPT-4o-mini	43.5	52.3	38.3	38.6	54.5	63.9	45.1	66.7	41.5	42.9
Claude-3-Sonnet	43.1	52.6	37.5	36.4	54.3	64.2	44.5	66.7	41.7	47.6
Claude-3.5-Sonnet	40.3	47.7	35.8	54.5	55.6	65.9	45.3	66.7	41.8	42.9

MMVP benchmark [65]. MMVP benchmark is an intentionally selected real image dataset which is claimed to be difficultly distinguished by MLLMs. We manually label 89 VQAs regarding camera-object relations. During evaluation, MLLMs' textual responses are evaluated by GPT-4o, which is prompted to judge if the responses correspond to the ground truth answers. We calculate the accuracy rate (%) of the MLLMs responses as the evaluation metric. We also include commercial SOTA models (e.g. GPT-4o and Claude-3.5-Sonnet) in the evaluation.

Quantitative Results. The quantitative results of response accuracy (%) are shown in Tab. 1. Baseline LLaVA models only provide one of several fixed options ("front" orientation, "top" viewpoint, etc.) regardless of the content of the input images. The accuracy of baselines are near random guess level (e.g. 33% or 12.5% depending on the option type). This weakness may result from the bias on LLaVA's visual instruction datasets. Since these datasets are based on real image datasets, most images are captured from a front facing direction, and from an above or side viewpoint.

All MLLMs finetuned by our synthetic generated dataset significantly improve the accuracy of recognition on camera-object relation. On *Ultimate3D* benchmark, fine-tuned LLaVA-1.6-13B model outperforms commercial SOTAs by average of 33.4% percentage. Moreover, the performance shows plausible generalization on real and synthetic image benchmarks. In MMVP benchmark, fine-tuned LLaVA-1.6-13B model also provides average of 19.23% higher accuracy than commercial SOTAs. Smaller LLaVA-1.6-7B model provides similarly competitive performances to its 13B version. This indicates the potential of using a compact model to distinguish the camera-object relations.

Moreover, Llama-3.2-Vision-11B obtains the largest accuracy improvement (average 60.46% ↑) on all three types of camera-object relation recognition tasks. The original released model refuses to answer most of the questions (e.g., response: "I'm not able to provide information about the individual in this image."). But our fine-tuned model provides the best performance on recognizing orientation compared to all alternative MLLMs.

Qualitative Results. We show the qualitative comparisons in Fig. 4. Commercial SOTA models (e.g. GPT-4o) and open-sourced models (e.g. LLaVA-1.6-13B) struggle with precise recognition of all three types of camera-object relations. For the two commercial models, their responses of predicting object orientation are near random regardless of the input images. We find a bias that both commercial models tend to underestimate viewpoint angle (e.g. mistake top view as horizontal), and overestimate camera-shot distance (e.g. mistake medium-shot as long-shot). The original LLaVA model has severe bias in its responses to all types of questions. Most of its answers are "Front Left" orientation, "Top" viewpoint, and "Close-up" camera-shots. As comparison, the LLaVA model fine-tuned by our *Ultimate3D* dataset provides plausible performance on all tasks across the synthetic and real sets of our benchmark. Additional comparisons are in Supplemental Sec. 8.

We hypothesize that both dataset bias and lack of camera-object visual instruction dataset cause the aforementioned behavior biases across various MLLMs. Without modifications to model structure or training scheme, our fine-tuned LLaVA-1.6-13B model significantly outperforms

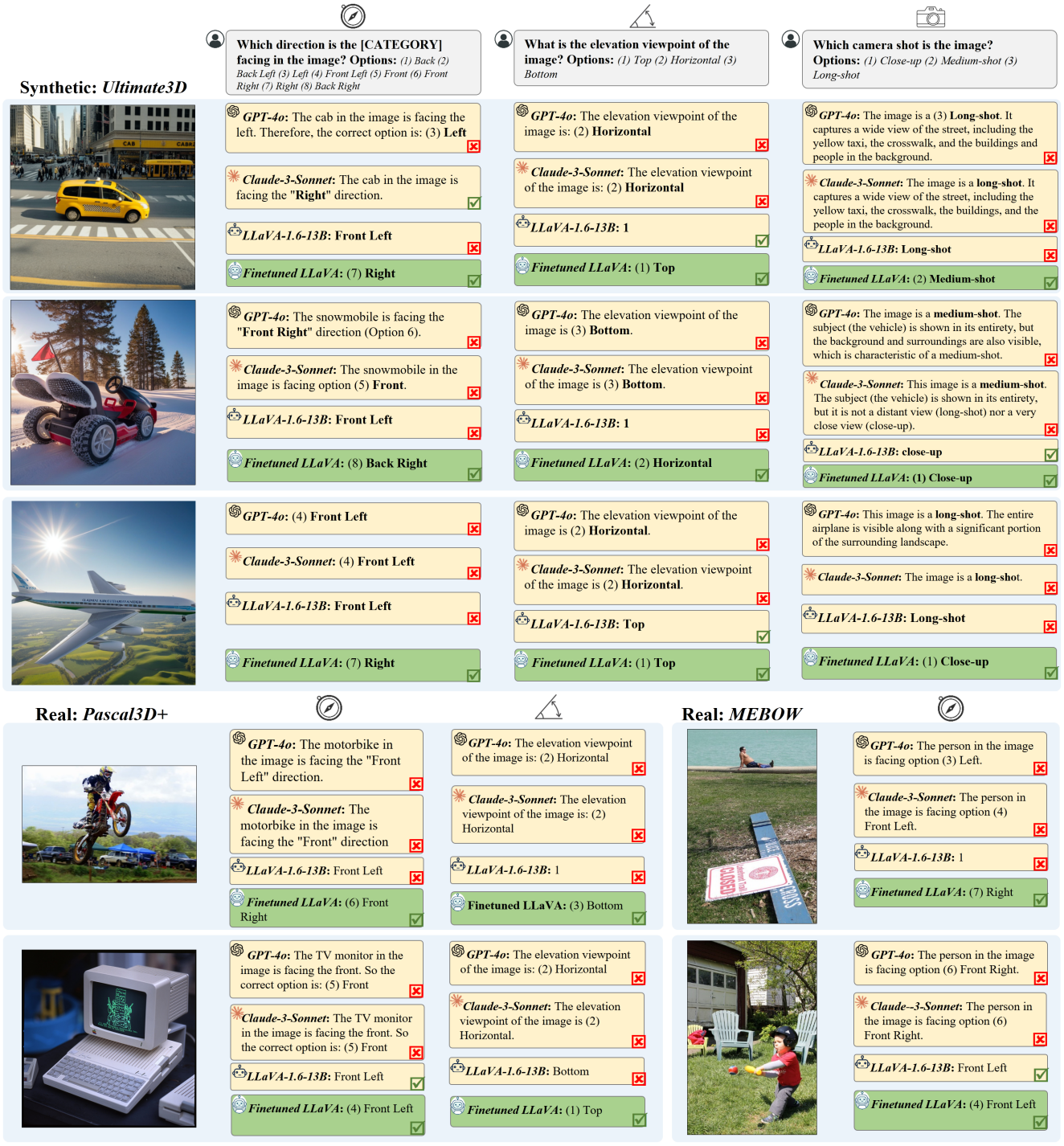

Figure 4. **Qualitative Results.** We show responses by GPT-4o, Claude-3-Sonnet, LLaVA-1.6-13B, and fine-tuned LLaVA-1.6-13B, on *Ultimate3D* benchmark. Each model is asked the questions (in gray boxes) regarding object orientation, camera viewpoint, and camera-shots type, together with the input images on the left. The model responses illustrate that the LLaVA-1.6-13B model fine-tuned by *Ultimate3D* can precisely recognize camera-object relations in both real and synthetic images. GPT-4o and Claude-3-Sonnet provide performance similar to random guess on object orientation, display a bias to underestimate viewpoint angle, and show a bias to overestimate camera-shot distance. The original LLaVA-1.6 model defaults to a few fixed answers regardless of the input images.

commercial models which have a large number of model parameters and consume huge amount of training data. This indicates that lack of dataset is the main reason of current MLLMs' weakness on camera-object relation recognition.

Limitations. The numerical prediction of camera-object relation β is more challenging for MLLMs compared to the categorical prediction we report. Preliminary tests show MLLMs can handle numerical prediction of camera-object

Table 2. **Generalization.** We finetune LLaVA-1.6-13B model using a mixture of both general VQA and camera-object relation VQA datasets. Without degradation on general VQA tasks, finetuned model (green) improves its ability on camera-object relation recognition and beat SOTA (i.e. GPT-4o) by 28.42% accuracy on *Ultimate3D* benchmark. $*$ denotes the average accuracy (%) across all 3 types of camera-object relation VQAs in Tab. 1.

MLLMs	VQAv2 [15]	*Ultimate3D**	*MMVP**
LLaVA-1.6-13B [40]	80.00	29.70	37.03
Ulti3D (100%)	78.77	**79.83**	**70.77**
Ulti3D (50%) + [40](50%)	**80.01**	74.72	65.80
GPT-4o	/	46.30	51.57
Claude-3.5-Sonnet	/	45.90	54.70

relation for a single category, but struggle with a hundred categories as in our dataset. Since our main contribution is the discovery of the dataset bottleneck and the generation pipeline, we leave addressing this limitation as future work.

4.4. Generalization of *Ultimate3D* to General VQA

To explore the generalization of 3D camera-object relation capturing with general VQA tasks, we fine-tune LLaVA-1.6-13B models by a half-half mixture of *Ultimate3D* (240K VQAs) and randomly sampled 240K VQAs from LLaVA-Instruct-665K [40]. Tab. 2 presents the dataset mixture finetuning performance on VQAv2 [15], our *Ultimate3D*, and *MMVP* benchmarks. The results demonstrate that incorporating *Ultimate3D* significantly enhances camera-object relation reasoning while maintaining general VQA performance.

4.5. Ablations on Image Generation Pipeline

In Fig. 5, we ablate our image generation pipeline and illustrate better quality than [45] by: (1) removing the visual priors; and (2) using alternative image generation backbones; (3) changing the group of visual priors given to the image generator. Compared to text-only SDXL and [45], we find that multiple visual priors may significantly improve the robustness of image generation under complex camera-object relation (especially for viewpoints of the object's back-side, and for 3D assets with low level-of-detail). In particular, the depth and Canny edge prior are crucial. However, we find that adding more priors (e.g., RGB, semantic masks, normal maps) does not explicitly improve quality. Moreover, our SDXL backbone outperforms other alternatives (i.e. SD-V1.5 [52]). Discussions in Suppl Sec. 5 and Sec. 6.

4.6. User Study on Dataset Quality

We investigate using text-image aligning metric and pose-estimation model prediction to evaluate quality of our Ultimate3D dataset but find those metrics to be deficient (Supp Sec. 4). Instead, we perform a comprehensive user study as a quality check of *Ultimate3D* dataset. As Supp Sec.7 shows, the user will see two images side-by-side (Supp Fig.4): an RGB image rendered by Blender (representing I_β), and a synthetic image generated by our method using I_β. Preserving the RGB image's 3D geometry and structure in the synthetic image is regarded as success. Our study includes 225 randomly sampled image pairs where each image is independently reviewed by 3 users. Overall success rate is **93.07**% (agreement rate among 3 users is 94.4%). In comparison, we perform another similar-objective user study on [45]. The overall success of this alternative method was only 55.11% (we explain the cause of this reduced performance in Sec. 4.5).

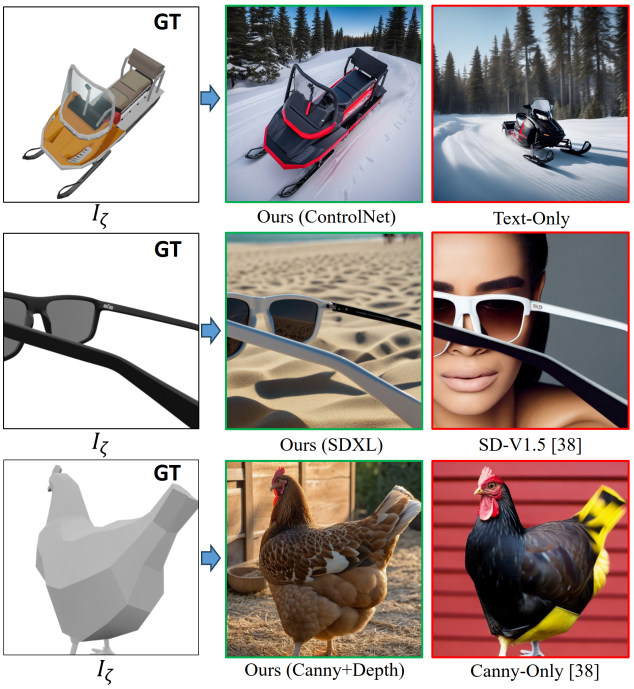

Figure 5. **Ablations on Image Generation Pipeline.** Our framework shows superior performance by using multi-ControlNet with SDXL backbones, and introducing both Canny edges and depth maps as visual priors. It consistently delivers better robustness and quality compared to text-only SDXL and [45]. When these visual priors are removed (top-right: removing all visual priors; bottom-right: removing depth prior), the generator often fails to produce good results and brings lower generation success rate (see Sec 4.6, [45] only achieves 55.11% success rate v.s. ours 93.07%)

5. Conclusion

In this work, we discover the dataset bottleneck of MLLMs for distinguishing camera-object relations. Given only 3D assets, our framework is able to generate unlimited image-text pairs of diverse camera-object relations with no distribution bias. We also provided *Ultimate3D* dataset and benchmark MLLM community. Moreover, fine-tuned open-sourced MLLMs using our dataset outperform SOTA commercial models by a significant margin. We plan to extend our framework for multi-object image generation with additional viewpoints, and improve numerical prediction capability on camera-object relation.

6. Acknowledgments

This work was performed during Liu He's internship at Amazon. We thank Jingjing Zheng, Tony Qi, Nan Qiao, Ke Zhang, and Yuyin Sun for their insightful discussions and valuable contributions to the formulation of the ideas in this work.

References

[1] Anthropic. Claude v3.0, 2024. https://www.anthropic.com/. 1

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[3] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 2

[4] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 2

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 2

[6] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024. 2

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 2

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 5

[10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[12] Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. *arXiv preprint arXiv:2311.01487*, 2023. 2

[13] Zilin Du, Haoxin Li, Xu Guo, and Boyang Li. Training on synthetic data beats real data in multimodal relation extraction. *arXiv preprint arXiv:2312.03025*, 2023. 2

[14] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 2

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2, 8

[16] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022. 2

[17] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806, 2024. 2

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2

[19] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 2

[20] Liu He and Daniel Aliaga. Globalmapper: Arbitrary-shaped urban layout generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 454–464, 2023. 3

[21] Liu He and Daniel Aliaga. Coho: Context-sensitive city-scale hierarchical urban layout generation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[22] Liu He, Yijuan Lu, John Corring, Dinei Florencio, and Cha Zhang. Diffusion-based document layout generation. In *International Conference on Document Analysis and Recognition*, pages 361–378. Springer, 2023.

[23] Liu He, Jie Shan, and Daniel Aliaga. Generative building feature estimation from satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[24] Liu He, Yizhi Song, Hejun Huang, Pinxin Liu, Yunlong Tang, Daniel Aliaga, and Xin Zhou. Kubrick: Multimodal agent collaborations for synthetic video generation. *arXiv preprint arXiv:2408.10453*, 2024. 3

[25] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[27] Hang Hua, Ziyun Zeng, Yizhi Song, Yunlong Tang, Liu He, Daniel Aliaga, Wei Xiong, and Jiebo Luo. Mmig-bench: Towards comprehensive and explainable evaluation of multi-modal image generation models. *arXiv preprint arXiv:2505.19415*, 2025. 3

[28] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[29] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2

[31] Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Christopher Pal, and Siva Reddy. Learning action and reasoning-centric image editing from videos and simulations. *arXiv preprint arXiv:2407.03471*, 2024. 3

[32] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1, 2

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2

[34] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024. 2

[35] Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*, 2023. 2

[36] Yuan-Hong Liao, Sven Elflein, Liu He, Laura Leal-Taixé, Yejin Choi, Sanja Fidler, and David Acuna. Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception. *arXiv preprint arXiv:2504.15362*, 2025. 3

[37] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 1, 2

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[39] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 2, 4

[40] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 2, 5, 8

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2, 4

[42] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194, 2019. 2

[43] Zheng Liu, Hao Liang, Wentao Xiong, Qinhan Yu, Conghui He, Bin Cui, and Wentao Zhang. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*, 2024. 2

[44] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2

[45] 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, 8

[46] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2

[47] Meta. Llama 3.2 vision, 2024. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. 1, 5

[48] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 2

[49] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[50] 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. 1

[51] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3, 4, 5

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 8

[53] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5

[54] 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 Processing Systems*, 35:25278–25294, 2022. 2

[55] Sahand Sharifzadeh, Christos Kaplanis, Shreya Pathak, Dharshan Kumaran, Anastasija Ilic, Jovana Mitrovic, Charles Blundell, and Andrea Banino. Synth2: Boosting visual-language models with synthetic captions and image embeddings. *arXiv preprint arXiv:2403.07750*, 2024. 2

[56] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 2

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[58] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023.

[59] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8048–8058, 2024.

[60] Yizhi Song, Liu He, Zhifei Zhang, Soo Ye Kim, He Zhang, Wei Xiong, Zhe Lin, Brian L Price, Scott Cohen, Jianming Zhang, et al. Refine-by-align: Reference-guided artifacts refinement through semantic alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

[61] Li Sun, Liu He, Shuyue Jia, Yangfan He, and Chenyu You. Docagent: An agentic framework for multi-modal long-context document understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17712–17727, 2025.

[62] Yunlong Tang, Jing Bi, Chao Huang, Susan Liang, Daiki Shimada, Hang Hua, Yunzhong Xiao, Yizhi Song, Pinxin Liu, Mingqian Feng, et al. Caption anything in video: Fine-grained object-centric captioning via spatiotemporal multimodal prompting. *arXiv preprint arXiv:2504.05541*, 2025.

[63] Yunlong Tang, Junjia Guo, Pinxin Liu, Zhiyuan Wang, Hang Hua, Jia-Xing Zhong, Yunzhong Xiao, Chao Huang, Luchuan Song, Susan Liang, et al. Generative ai for cel-animation: A survey. *arXiv preprint arXiv:2501.06250*, 2025. 3

[64] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1, 2, 3

[65] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 2, 5, 6

[66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[67] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023. 2

[68] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18359–18369, 2023. 3

[69] 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. 5

[70] 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. 5

[71] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn

of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 2

[72] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2

[73] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2

[74] Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7920–7930, 2025. 3

[75] Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. *arXiv preprint arXiv:2509.21309*, 2025. 3

[76] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 2

[77] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[78] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 4

[79] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 2

[80] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 2

[81] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 2