This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Is 'Right' Right? Enhancing Object Orientation Understanding in Multimodal Large Language Models through Egocentric Instruction Tuning

Ji Hyeok Jung[1] Eun Tae Kim[1] Seoyeon Kim[1] Joo Ho Lee[1] Bumsoo Kim[2,*] Buru Chang[3,*]
[1]Sogang University [2]Chungang University [3]Korea University
{ji9759,untae0122,ksy02031,jhleecs}@sogang.ac.kr bumsoo@cau.ac.kr buru_chang@korea.ac.kr

Abstract

Multimodal large language models (MLLMs) act as essential interfaces, connecting humans with AI technologies in multimodal applications. However, current MLLMs face challenges in accurately interpreting object orientation in images due to inconsistent orientation annotations in training data, hindering the development of a coherent orientation understanding. To overcome this, we propose egocentric instruction tuning, which aligns MLLMs' orientation understanding with the user's perspective, based on a consistent annotation standard derived from the user's egocentric viewpoint. We first generate egocentric instruction data that leverages MLLMs' ability to recognize object details and applies prior knowledge for orientation understanding. Using this data, we perform instruction tuning to enhance the model's capability for accurate orientation interpretation. In addition, we introduce EgoOrientBench, a benchmark that evaluates MLLMs' orientation understanding across three tasks using images collected from diverse domains. Experimental results on this benchmark show that egocentric instruction tuning significantly improves orientation understanding without compromising overall MLLM performance. The instruction data and benchmark dataset are available on our project page at https://github.com/jhCOR/EgoOrientBench.

1. Introduction

Multimodal large language models (MLLMs) [9, 30, 52] can serve as interfaces connecting humans with AI technologies designed to perform multimodal tasks that require both image and text understanding. For instance, MLLMs are useful for operating autonomous systems [40, 46], manipulating robots [18, 22], and facilitating communication with AI assistants on AR devices [33, 35, 45]. To ensure consistent interaction, it is essential for MLLMs to exhibit behaviors that align accurately with human intentions.

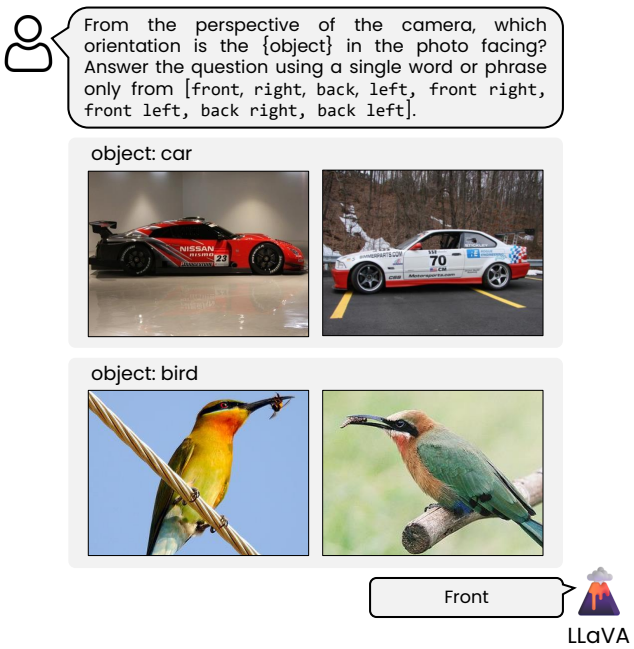


Figure 1. Examples of LLaVA's [30] responses to prompts asking about the orientation of objects in an image. Current MLLMs show a significant lack of understanding when interpreting the orientation of given objects.

However, MLLMs' limited understanding of object orientation complicates effective communication. As shown in Figure 1, current MLLMs struggle to accurately interpret the orientation an object is facing. This limitation can lead to errors in autonomous vehicle navigation, incorrect robot operations, and even severe accidents, creating a significant barrier to deploying MLLMs in real-world applications.

In this study, we identify inconsistent annotations in training data—without standardized guidelines—as a major factor hindering accurate orientation understanding. Unlike humans, who understand object orientation through an egocentric reference frame based on their physical bodies [20, 43, 51], MLLMs, lacking a physical body, rely

*Corresponding authors.



Figure 2. Examples of inconsistent annotations for object orientation. In image-text pairs used for training MLLMs, such as those in MSCOCO [29] and LAION-5B [39], annotations for object orientation lack consistency. For instance, (a) objects facing different orientations may be annotated as facing the same orientation, or conversely, (b) objects facing the same orientation may be annotated differently. This variation arises because, without a standardized guideline for object orientation, annotations can vary depending on individual perspectives or cultural backgrounds [26]. Our study aims to improve MLLM's understanding of object orientation by aligning it with the user's egocentric perspective through instruction tuning.

solely on training data to develop orientation understanding. However, as shown in Figure 2, datasets used to train MLLMs like MS-COCO [29] and LAION-5B [39] contain inconsistent annotations for object orientation, sometimes based on the observer's perspective and other times on the object's perspective. Such inconsistent annotations for objects make it difficult for MLLMs to understand object orientation, leading to behaviors misaligned with user intentions. Since annotations can vary depending on individual or cultural backgrounds [26], annotation inconsistencies are inevitable without a consistent annotation standard that considers real-world use cases for MLLMs.

To address this issue, we propose *Egocentric Instruction Tuning*, which aligns MLLMs' understanding of object direction with the user's perspective. Considering practical application scenarios where users and MLLMs interact from a unified viewpoint, this tuning method adjusts the MLLM's orientation understanding to align with the user's egocentric perspective. To implement this, we manually annotate object orientation based on an egocentric perspective using ImageNet data [10], then generate egocentric instruction data from these annotations. This process leverages MLLM's ability to recognize image details and the LLM's prior knowledge to associate these details with object orientation. By applying instruction tuning with this data, we improve alignment with human perception of object orientation without compromising general performance.

Furthermore, we introduce *EgoOrientBench* (Egocentric Object Orientation Understanding Benchmark), a benchmark to evaluate MLLM object orientation comprehension. Existing datasets [42] for assessing this capability are limited to small collections with only a few dozen samples, restricting comprehensive evaluation. Our benchmark includes three tasks of varying complexity to enable a thorough assessment of object orientation comprehension. Evaluation data is collected from diverse domains—including photographs, painting, drawing, and 3D rendering—to reflect a wide range of application use cases. Experimental results on this benchmark demonstrate that aligning object orientation understanding in MLLMs with the user perspective through egocentric instruction tuning enhances applicability in real-world scenarios.

Contributions. Our contributions can be summarized as follows: (1) We introduce a new task to enhance and evaluate MLLMs' understanding of object orientation from the user's egocentric perspective. (2) We propose *egocentric instruction tuning*, a method that aligns MLLM's understanding of object orientation with the user's egocentric perspective, and provide the necessary data for this approach. This method enhances orientation understanding without compromising general performance. (3) We introduce *EgoOrientBench*, a benchmark designed for comprehensive evaluation of object orientation understanding in MLLMs.

2. Related Work

2.1. Multimodal Large Language Model

Recently, numerous studies have concentrated on developing MLLMs that can effectively interpret both visual and textual data. For example, Flamingo [2] leverages visual and textual inputs as prompts, achieving notable few-shot performance in visual question answering. Models like

LLaVA [30] and MiniGPT4-v2 [52] apply visual instruction tuning to align MLLM behavior more closely with user intentions. Moreover, methods such as VisionLLM [44], KOSMOS-2 [37], and Qwen-VL [4] have demonstrated superior performance in visual grounding tasks. InternVL [7] underscores the significance of scaling the visual encoder in proportion to the LLM's size, while PaLM-E [14] and EmbodiedGPT [33] reveal the promising potential of MLLMs for embodied applications. Additionally, commercial APIs, including GPT-4 [1], Claude 3.5 [3], and Gemini-1.5 [41], are actively utilizing MLLMs across a range of applications.

Nevertheless, previous research [19, 42] has identified that current MLLMs struggle with accurately understanding object orientation. In this study, we highlight that this limitation arises from the inconsistent annotations regarding object orientation in the image-text pair data used for MLLM training. To address this issue, we focus on developing an egocentric instruction tuning method that aligns MLLMs' understanding of object orientation with human perception, thereby enhancing orientation understanding in user-centered applications.

2.2. Object Orientation Understanding

Research on object orientation understanding is actively pursued across various fields. Pose estimation [11, 34] is a fundamental task in interpreting object orientation in images. In autonomous driving, for instance, understanding pedestrian body orientation aids in predicting movement paths and reducing accident risk [12, 15, 17]. Recent studies have explored enhancing object orientation comprehension by incorporating viewpoint information. For example, DREAMFUSION [38] uses orientation data from multiple viewpoints to create enriched object descriptions, while EgoExoLearn [24] focuses on harmonizing orientation interpretation across varying perspectives.

More recently, MMVP [42] has underscored the need for object orientation understanding within MLLMs. To evaluate encoder representations, MMVP introduces a simple evaluation method and dataset that measure MLLMs' ability to interpret object orientation. However, this approach is limited by a small sample size of real-image data, which restricts comprehensive assessment across diverse image domains and various orientation complexities. To address this limitation, we propose a large-scale benchmark specifically designed to evaluate MLLMs' object orientation understanding across a broad range of domains and conditions, facilitating a more robust analysis of MLLM capabilities.

3. Egocentric Instruction Tuning

3.1. Egocentric Annotation

Motivation. Inconsistent annotations hinder MLLMs' ability to develop an understanding of object orientation. Therefore, consistent annotations should be used in MLLM training, which requires a standardized annotation guideline. Meanwhile, with the rise of embodied AI, numerous studies [5, 21, 48] have explored using AI technologies from a user's egocentric perspective. For instance, over ten egocentric video datasets, such as EgoExoLearn [24], EgoCoT [33], and EgoSchema [32], have been introduced in the past year. Motivated by this trend, we standardize annotations based on the user's egocentric viewpoint to address the issue of inconsistent object orientation annotations. This standardization enhances the applicability of MLLMs in real-world, user-centered applications.

Orientation Classes. We construct the egocentric instruction data using a consistent annotation rule. This rule categorizes object orientation into eight distinct egocentric classes (see Figure 4):

- ***Front***: the object faces the user (or camera).
- ***Back***: the object is turned away, facing the opposite direction from the user (or camera).
- ***Left***: the object is oriented to the left of the user.
- ***Right***: the object is oriented to the right of the user.
- ***Front-Left***, ***Front-Right***: the object faces the user but is angled toward the left (or right).
- ***Back-Left***, ***Back-Right***: the object faces away from the user but is angled toward the left (or right).

This classification scheme provides a structured framework to interpret object orientation consistently from a user-centered perspective.

3.2. Egocentric Instruction Data

To create the egocentric instruction dataset, we begin by manually annotating object orientations using the ImageNet data [10]. We then generate LLaVA-style instruction data [30] leveraging MLLM's ability to recognize image details and the LLM's prior knowledge to associate these details with object orientation. Using this dataset, we perform instruction tuning to enhance the MLLMs' ability to interpret object orientation while maintaining their general response generation capabilities.

Manual Data Collection. Among the datasets used for training MLLMs, none provide egocentric annotations for object orientation. To address this gap, we initiate the first effort to manually annotate object orientation in an egocentric context, creating a unique dataset specifically designed to enhance MLLMs' understanding of object orientation. We focus initially on single-object scenarios to establish a foundation for further exploration and development. Using ImageNet data [10] and the annotation scheme detailed in Section 3.1, we classify objects into eight distinct orientations, ensuring clarity and minimizing ambiguity in the dataset. Dataset statistics are summarized in Appendix A.2.

Instruction Data Generation. Using manually annotated orientation information, we create egocentric instruction

Prompt:
As you are directional aware LVLM and observer that have no linkage of the object, closely take a look and answer the question {Image Tag Here}.



Response Type 1: QA with Detail Recognition:
Question: From the camera perspective, is the common *agama* facing away from the camera or facing towards the camera? first describe what you can find from the object in the image, then based on that, answer the orientation of the object.
Answer: (what i find): tail, legs, body, head, and neck. (answer the orientation): according to (what i find), the agama is facing right.



Response Type 2: QA with Detail Recognition and LLM Prior Knowledge:
Question: Describe in detail the features in the image that indicate the orientation of the *sports car*.
Answer: The sports car is parked on a dirt road, and the front end is facing the camera. the headlights are on the right side of the image, and the front grill is visible. The car's orientation is definitively towards the front right.



Response Type 3: QA for Object Alignment
Question: Which orientation would the *tractor* have to turn to face the camera directly with the least angle of rotation?
Answer: The tractor is turned to the right, so it needs to turn clockwise to face the camera directly.

Figure 3. An example of egocentric instruction data designed to enhance MLLMs' understanding of object orientation. This data leverages the model's intrinsic ability to recognize object details (Response Type 1) and the LLM's prior knowledge to link these details to specific orientations (Response Type 2). Furthermore, by engaging in object alignment tasks that require understanding the relationships between different orientations (Response Type 3), MLLMs' comprehension of object orientation is further improved.

tuning data inspired by LLaVA data [30] to align object orientation understanding with the user's egocentric perspective while preserving MLLMs' general response generation capabilities. Instead of directly answering an object's orientation, we design three response types to enhance orientation understanding by leveraging the model's ability to recognize image details and the LLM's prior knowledge:

- **Response Type 1**: Utilizes MLLM's detail recognition abilities. The model identifies specific image details, such as a car's taillights or an elephant's tail, that provide clues about orientation, guiding MLLM to interpret the object's orientation more accurately.
- **Response Type 2**: Draws on the LLM's prior knowledge essential for understanding orientation, instructing the model to describe orientation based on key details. For example, the LLM knows that a human face's orientation depends on the nose position, while a lizard's orientation relies on the head-tail alignment. This type links prior knowledge with image details, enabling MLLM to describe the object and conclude its orientation.
- **Response Type 3**: Adapts from robot manipulation tasks to improve MLLM's understanding of orientation relationships. Aligning an object to a target orientation requires grasping the relation between its current and target orientations. For instance, rotating an object facing front-left counterclockwise aligns it to front. This type enhances MLLM's understanding of orientation relationships from an egocentric perspective.

In Section 5.3, we will conduct ablation tests to validate the contribution of each response type to improved object orientation understanding.

3.3. Instruction Tuning

We conduct instruction tuning on the MLLM using the previously generated egocentric instruction data. During this process, we keep the visual encoder frozen and apply supervised fine-tuning to both the LLM and the bridge layer. To increase training efficiency, we employ LoRA [23] for parameter-efficient fine-tuning.

4. *EgoOrientBench*: Egocentric Orientation Understanding Benchmark

4.1. Data Collection.

MLLMs, known for their versatile capabilities, can be applied across various domains. To assess MLLMs' ability to recognize object orientation in diverse domains, we introduce *EgoOrientBench*, a large-scale benchmark dataset designed for orientation understanding from a user-centered, egocentric perspective. The dataset is constructed from five distinct image datasets, spanning various domains such as drawings, paintings (from PACS and DomainNet), and 3D-rendered images (from OmniObject3D).

- **ImageNet** [10]: Provides real images of single objects across 1,000 object classes.
- **D_3** [19]: Offers a small-scale data designed for a comprehensive assessment of MLLM's object understanding capabilities, including several dozen examples that evalu-

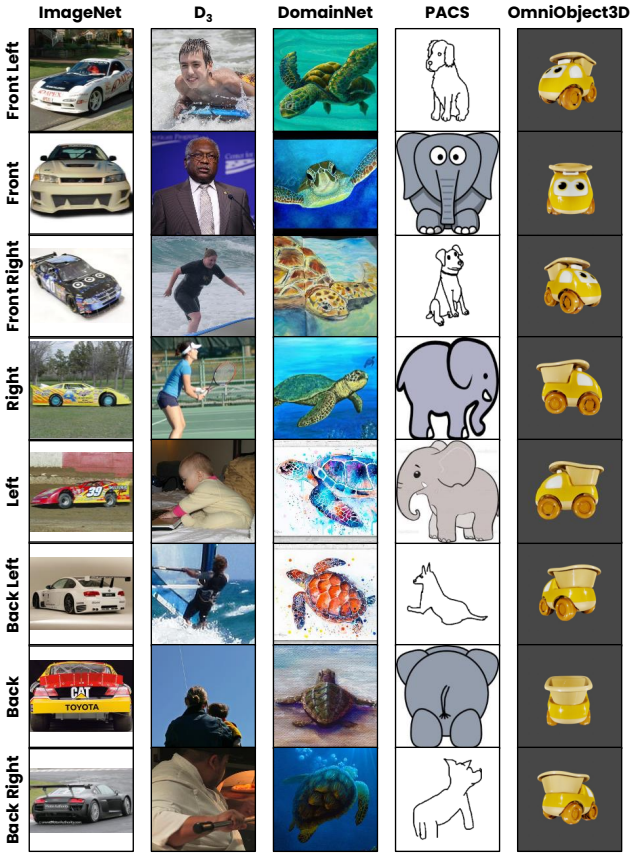


Figure 4. Our benchmark data examples. Our benchmark consists of data collected from various image domains to assess the applicability of MLLMs in terms of orientation understanding. The collected data is annotated across eight orientation classes.

ate object orientation in real images.

- **DomainNet** [36]: Includes cross-domain evaluation images, featuring single-object representations in sketches, paintings, and clip art.
- **PACS** [27]: Contains images from multiple domains, including real photos, art paintings, cartoons, and sketches.
- **OmniObject3D** [47]: Supplies mesh data for 3D rendering, enabling the capture of object orientations by adjusting camera angles.

We annotate the data across these sources according to the eight orientation classes outlined in Section 3.1. For the ImageNet dataset, we collected 50 images per class across the eight orientations. In the D_3, DomainNet, and PACS datasets, images for the *Back*, *Back-Left*, and *Back-Right* orientations are sparse, resulting in imbalanced annotations. OmniObject3D, which allows rendering of an object in all eight orientations by adjusting the camera angle, enable balanced annotation across orientations. Figure 4 provides examples of this benchmark evaluation data, and detailed statistics are presented in Appendix A.2.

4.2. Task Description.

We design three tasks to assess MLLMs' object orientation understanding at various levels

- ***Choose***: A simple classification task where the model selects an object's orientation from eight multiple-choice options to assess basic orientation understanding.
- ***Verify***: A binary classification task where the model determines whether an object's orientation matches a given query, instead of choosing from set options.
- ***Freeform***: Unlike the previous tasks, this task involves generating a descriptive response for an object's orientation, rather than selecting from options. GPT-4 then assesses the alignment between the generated description and the true orientation.

Each task targets specific aspects of orientation understanding, while offsetting each other's limitations. For example, *Choose* is effective for basic orientation comprehension but limited by the division of continuous orientation into discrete classes. To overcome this, *Verify* tests the model's ability to match the given orientation in the query to an object in an image. Finally, since descriptive vocabulary can vary, *Freeform* provides flexible evaluation by generating descriptive orientation responses, matched to the true answer using GPT-4. Task prompts of EgoOrientBench are detailed in Appendix A.3.

5. Experiments

5.1. Experimental Setup

Backbone Models. To measure the current object orientation understanding of MLLMs and verify the effectiveness of our egocentric instruction tuning, we evaluate the following three MLLMs in EgoOrientBench: LLaVA 1.5 (w/ Vicuna-7B) [30], mPLUG-Owl2 (w/ LLaMA2-7B) [49], and InternVL2 (4B) [7]. Additionally, we also conduct evaluations for commercial MLLMs, GPT-4o (`gpt-4o-2024-08-06`) [1] and Gemini-1.5 (`Flash-8B`) [41] using their API.

Benchmarks. To verify the effectiveness of egocentric instruction tuning, we conduct experiments in ***EgoOrientBench*** to assess MLLMs' understanding of object orientation. Additionally, to determine whether egocentric instruction tuning enhances object orientation understanding without compromising general response generation, we evaluate the models' generalization capabilities using four benchmarks: ***MME*** [16], ***MMStar*** [6], ***MMMU*** [50], and ***POPE*** [28].

Implementation Details. All models are trained on a single A6000 48GB GPU for 3 epochs. To maintain general response generation capabilities, the visual encoder remains frozen across all models. In our implementation, we train the LLaVA-1.5, mPLUG-Owl2, and InternVL2 models following the hyperparameters specified in the original papers.

Method		Choose						Verify						Freeform					
Backbone	Method	IN	D_3	DN	PA	3D	Avg.	IN	D_3	DN	PA	3D	Avg.	IN	D_3	DN	PA	3D	Avg.
-	*Popular*	12.5	25.0	27.7	30.7	12.5	20.3	50.0	50.0	50.0	50.0	50.0	50.0	-					
LLaVA-1.5 (7B)	Zero-shot	12.2	22.9	24.0	23.9	12.5	17.9	52.2	49.3	51.3	51.2	52.3	51.8	14.7	15.1	21.3	20.5	15.5	17.9
	Ours	**42.4**	**37.2**	**42.4**	**40.8**	**25.2**	**33.7**	**62.2**	**56.0**	**61.6**	**58.3**	**56.7**	**58.3**	**65.5**	**50.0**	**72.3**	**67.2**	**43.3**	**56.5**
mPLUG-Owl2 (7B)	Zero-shot	14.2	23.4	18.4	15.9	12.8	15.0	54.4	49.5	50.2	46.4	50.9	49.8	19.3	28.3	26.6	26.6	28.3	20.7
	Ours	**34.5**	**34.0**	**33.0**	**37.1**	**21.2**	**28.5**	**67.1**	**62.8**	**64.2**	**65.4**	**57.6**	**61.5**	**41.5**	**40.1**	**42.5**	**48.1**	**28.4**	**37.1**
InternVL2 (4B)	Zero-shot	15.4	15.3	21.1	15.7	15.2	16.5	54.1	56.6	63.0	59.5	54.9	57.7	22.0	39.5	53.1	52.3	28.9	39.6
	Ours	**18.7**	**30.9**	**42.0**	**50.3**	**18.3**	**31.4**	**58.8**	**58.5**	**68.2**	**67.0**	**56.0**	**61.4**	**41.2**	**43.4**	**66.8**	**66.1**	**31.6**	**48.2**
API	Gemini-1.5	24.5	25.7	29.3	34.8	30.0	30.7	55.0	53.0	56.1	61.5	56.5	57.5	36.2	51.3	55.0	61.3	49.3	52.9
	GPT-4o	36.0	44.7	47.5	55.8	31.1	41.1	75.7	77.0	79.8	77.8	73.8	76.2	58.3	63.2	86.8	87.6	59.8	72.3

Table 1. EgoOrientBench results (%). We report the results for three tasks (Choose, Verify, Freeform) on five datasets, ImageNet (IN) [10], D_3 [19], DomainNet (DN) [36], PACS (PA) [27], and OmniObject3D (3D) [47]. *Popular* indicates the performance achieved by uniformly responding with the most major class in each task, serving as the most basic baseline performance on imbalanced data.

All models are trained using the AdamW optimizer [31]. The learning rates are set as follows: $2e$-4 for LLaVA-1.5, $1e$-4 for mPLUG-Owl2, and $4e$-5 for InternVL2. A warm-up ratio of 0.3 and a cosine learning rate scheduler are applied to all models. Additionally, weight decay of 0.05 is applied exclusively to InternVL2. We use LoRA [23] for parameter-efficient fine-tuning, with settings of r=128 and α=256 for LLaVA-1.5 and mPLUG-Owl2, and r=16 and α=32 for InternVL2. Batch sizes are set according to each model's memory needs: LLaVA-1.5 uses a batch size of 4 with gradient accumulation of 4, mPLUG-Owl2 uses a batch size of 2 with gradient accumulation of 16, and MiniGPT-V2 uses a batch size of 1. During inference, tasks are validated one at a time without batching.

5.2. Experimental Result

Results on EgoOrientBench. Table 1 presents the evaluation results on EgoOrientBench, showing that our egocentric instruction tuning significantly enhances MLLMs' understanding of object orientation across all tasks. Specifically, all zero-shot MLLMs perform worse on the *Choose* task compared to the *Popular* method, which uses the most common orientation in each dataset as the default response. This performance gap stems from a strong orientation bias in current MLLMs on the *Choose* task, as analyzed in Section 5.3. The improvements achieved with our method suggest that it effectively mitigates this bias, resulting in better orientation comprehension. Although the commercial MLLM GPT-4o achieves the highest overall performance, our improved versions of LLaVA-1.5 and InternVL2 outperform another commercial model, Gemini-1.5, in all tasks.

Note that this enhancement occurs even with the visual encoder frozen, without additional training. While previous studies have attributed MLLMs' limitations in object orientation understanding to the visual encoder's inability to represent orientation information [19, 42], our benchmark experiments demonstrate that substantial improvements can be achieved by aligning MLLM orientation interpretation with the user's egocentric perspective.

Method		MME	MMStar	MMMU	POPE
Backbone	Method				
LLaVA 1.5	Zero-shot	1792.8	34.67	35.11	82.03
	Ours	1752.8	35.87	34.44	88.36
mPLUG-Owl2	Zero-shot	1706.3	34.33	37.55	86.16
	Ours	1727.3	35.27	38.55	85.60
InternVL2-4B	Zero-shot	2088.7	54.26	47.22	85.91
	Ours	2045.9	53.13	48.00	85.56

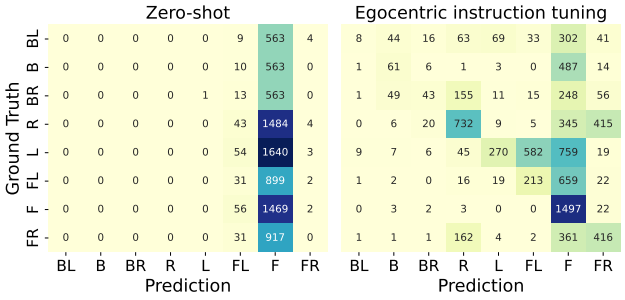
Table 2. Experimental results evaluating the general performance of MLLMs on the MME [16], MMStar [6], MMMU [50], and POPE [28] before and after egocentric instruction tuning.

Results on general benchmarks. Table 2 presents the experimental results on the general performance benchmarks. The results indicate that performance remains stable after applying our egocentric instruction tuning, with no significant degradation compared to pre-tuning levels. Overall, the three models exhibit largely comparable performance across the four benchmarks. In particular, LLaVA 1.5 shows a significant improvement in *POPE*, while mPLUG-Owl2 demonstrates considerable gains in *MME*. Similarly, InternVL2 maintains stable or slightly improved performance without noticeable degradation across most benchmarks.

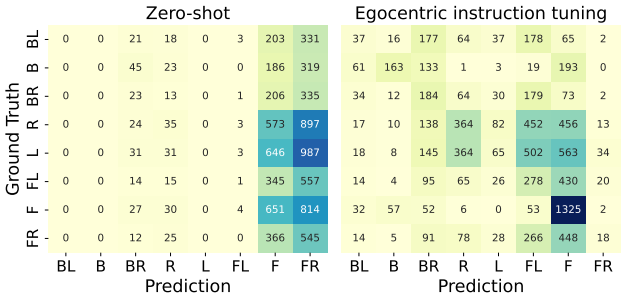
These findings demonstrate that our instruction tuning effectively aligns MLLM's object orientation understanding with the user's perspective while maintaining the model's overall response generation capabilities.

5.3. Analysis

Confusion Matrix. We analyze error cases before and after egocentric instruction tuning by examining the confusion matrix for the *Choose* task. As shown in Figure 5, LLaVA and mPLUG-Owl2 produce zero-shot predictions that are heavily biased toward the specific orientations, particularly the *Front* (F) and *Front Right* (FR) classes, indicating a significant deficiency in object orientation under-



(a) LLaVA



(b) mPLUG-Owl2

Figure 5. Confusion matrix for the *Choose* task with LLaVA and mPLUG-Owl2. Zero-shot MLLMs show extreme bias toward the *Front* class (F) (or the *Front Right* (FR) class). Our proposed egocentric instruction tuning relieves this bias and enhances the understanding of object orientation.

standing in publicly available MLLMs. We also observe that this bias is mitigated through our proposed egocentric instruction tuning. Notably, there are several acceptable error cases, such as predicting objects in the *Left* (L) class as *Front Left* (FL) or interpreting *Front Right* (FR) objects as *Front* (F) or *Right* (R) classes. This confusion matrix analysis is consistent with the benchmark results in Table 1, further confirming that our tuning method effectively enhances object orientation understanding. We want to emphasize that our EgoOrientBench provides a foundation for analyzing MLLM behaviors related to object orientation, supporting future research efforts to enhance MLLM orientation comprehension. The confusion matrix analysis for InternVL2 is provided in Appendix C.1.

Ablation Test. Table 3 shows the ablation test results for the three response types in our egocentric instruction data. Using all three response types yields the highest performance, while removing any type results in a decrease in task accuracy, indicating that each type plays a role in enhancing MLLM's object orientation understanding. Notably, combining Type 1 and Type 2 leads to significant improvements across all tasks, suggesting a strong synergy between them. Adding Type 3 further boosts performance, showing that it contributes to orientation understanding in a way that complements Types 1 and 2. The ablation results for mPLUG-Owl2 and InternVL2 are presented in Appendix C.2.

Response Types			Choose	Verify	Freeform
Type 1	Type 2	Type 3			
✓	✗	✗	18.3	54.7	32.4
✗	✓	✗	18.3	54.5	38.6
✗	✗	✓	30.0	50.3	46.1
✓	✓	✗	23.3	59.3	43.4
✓	✗	✓	27.7	56.4	46.6
✗	✓	✓	29.7	54.9	54.6
✓	✓	✓	33.7	58.3	56.5

Table 3. Ablation test with LLaVA-1.5. Each response type contributes to performance improvements across all three tasks.

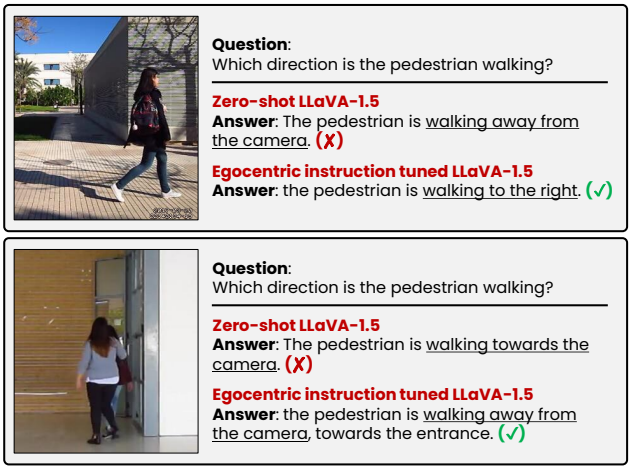


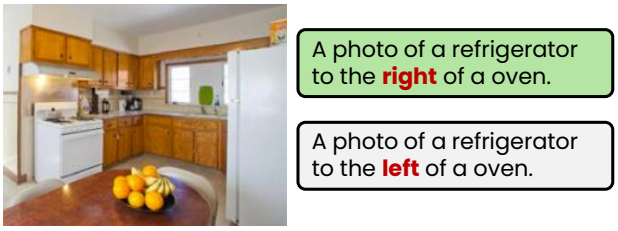
Figure 6. Pedestrian walking direction prediction results. These examples show that MLLMs with improved object orientation understanding, achieved through our egocentric instruction tuning, can be effectively applied to predicting pedestrian direction.

6. Discussion

6.1. Applicability in Real-world Scenarios

Pedestrian Walking Direction Prediction. If the direction of a pedestrian is accurately determined, their path can be correctly predicted, which helps reduce traffic accidents [12, 15, 17]. To explore the potential of using MLLMs in pedestrian walking direction prediction, we perform a qualitative analysis using pedestrian walking direction prediction data [13].

As shown in Figure 6, when prompt about a pedestrian's direction, zero-shot LLaVA-1.5 produces responses that do not match the pedestrian's actual direction. On the other hand, with our egocentric instruction tuning, the model generates responses that accurately align with the correct direction. Additionally, our method not only accurately describes the orientation of objects (*e.g.*, 'walking away from the camera') but also correctly represents the relationship



Backbone	Accuracy	
	Zero-shot	Ours
LLaVA-1.5	47.5	**62.5**
mPLUG-Owl2	25.7	**40.0**
InternVL	53.9	**57.3**

Table 4. Example and evaluation results for the spatial reasoning task on prepositional relationships between objects. The experimental results demonstrate that MLLMs with enhanced object orientation understanding through Egocentric Instruction Tuning also show improved spatial reasoning abilities for interpreting prepositional relationships between objects.

between the object and its surroundings (*e.g.*, 'towards the entrance'). Although our egocentric instruction tuning has not been specifically trained on pedestrian direction prediction tasks or images from that domain, it shows improved performance in this task, highlighting its practical utility.

Spatial Reasoning. Recently, spatial reasoning research [8, 25] aimed at understanding spatial relationships between objects using MLLMs has been actively conducted. Here, we present that aligning object orientation understanding with the user's egocentric perspective, thereby enhancing the ability to interpret object orientation itself, can improve MLLM's spatial reasoning capabilities. To support this, we evaluate Zero-shot MLLMs and MLLMs enhanced with our egocentric instruction tuning using the COCO-Spatial evaluation dataset from prior research [25].

As shown in the Table 4, our egocentric instruction tuning effectively improves spatial reasoning abilities for understanding preposition relationships between objects. Specifically, accuracy improved by 15.0%p, 14.3%p, and 3.4%p in LLaVA-1.5, mPLUG-Owl2, and InternVL2, respectively. This indicates that our approach can enhance MLLM performance in user-centered applications that require recognizing object positions, understanding orientations, and interpreting their spatial relationships.

6.2. Limitations

Figure 7 presents examples of error cases in LLaVA-1.5 after applying our egocentric instruction tuning. In these examples, the zebra and giraffe exhibit different orientations for their bodies and heads. Unlike inanimate objects such as cars or tractors, living beings like zebras and humans can

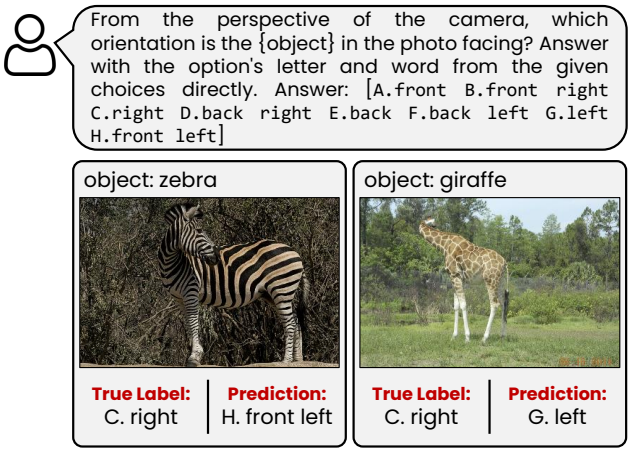


Figure 7. An example of an error case in our method. Incorrect predictions occur in complex situations where the body and head face different orientations, as seen with the zebra and giraffe.

adopt complex, independent orientations. These complexities pose challenges for our Response Type 2 data, which links object detail recognition with orientation understanding—for instance, when the head and tail suggest different orientations. Similarly, they complicate Response Type 3 data, which depends on interpreting orientation relationships for alignment tasks. To address these challenges, future work will explore approaches beyond the current discrete orientation categories.

7. Conclusion

In this study, we identify inconsistent object orientation annotations as a primary factor hindering MLLM's orientation understanding and propose an *Egocentric Instruction Tuning* method to address this by aligning MLLM's object orientation comprehension with the user's egocentric perspective. Additionally, we introduce *EgoOrientBench*, a benchmark designed to comprehensively evaluate MLLM's orientation understanding across three tasks and five image datasets. Experimental results on this benchmark, as well as on the general performance benchmarks, demonstrate that our egocentric instruction tuning effectively enhances object orientation understanding without compromising general response generation capabilities. Our analysis also reveals that current MLLMs exhibit a strong bias toward specific orientations, which our instruction tuning mitigates. Finally, we apply the enhanced MLLM, trained with egocentric instruction tuning, to pedestrian walking direction prediction and spatial reasoning tasks involving spatial relationships between objects. This demonstrates the practical benefits of enhanced orientation understanding for real-world applications. We believe our instruction tuning method and benchmark will significantly contribute to future research on MLLM's object orientation understanding.

Acknowledgment

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICT Creative Consilience Program grant funded by the Korea government(MSIT)(IITP-2025-RS-2020-II201819) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00212828, RS-2024-00350430).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[3] AI Anthropic. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 2024. 3

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 3

[5] Feiyu Chen, Cong Xu, Qi Jia, Yihua Wang, Yuhan Liu, Haotian Zhang, and Endong Wang. Egocentric vehicle dense video captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 137–146, 2024. 3

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 5, 6

[7] 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. 3, 5, A2

[8] Iulia Comșa and Srini Narayanan. A benchmark for reasoning with spatial prepositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16328–16335, 2023. 8

[9] 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

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3, 4, 6, A1

[11] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022. 3

[12] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011. 3, 7

[13] Alex Dominguez-Sanchez, Miguel Cazorla, and Sergio Orts-Escolano. Pedestrian movement direction recognition using convolutional neural networks. *IEEE transactions on intelligent transportation systems*, 18(12):3540–3548, 2017. 7

[14] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488. PMLR, 2023. 3

[15] Markus Enzweiler and Dariu M Gavrila. Integrated pedestrian classification and orientation estimation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 982–989. IEEE, 2010. 3, 7

[16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 5, 6

[17] Tarak Gandhi and Mohan Manubhai Trivedi. Image based estimation of pedestrian orientation for improving path prediction. In *2008 IEEE Intelligent Vehicles Symposium*, pages 506–511. IEEE, 2008. 3, 7

[18] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *International Conference on Robotics and Automation (ICRA)*, 2024. 1

[19] Manu Gaur, Makarand Tapaswi, et al. Detect, describe, discriminate: Moving beyond vqa for mllm evaluation. *arXiv preprint arXiv:2409.15125*, 2024. 3, 4, 6, A1

[20] Klaus Gramann, Julie Onton, Davide Riccobon, Hermann J Mueller, Stanislav Bardins, and Scott Makeig. Human brain dynamics accompanying use of egocentric and allocentric reference frames during navigation. *Journal of cognitive neuroscience*, 22(12):2836–2849, 2010. 1

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 3

[22] Tianrui Guan, Yurou Yang, Harry Cheng, Muyuan Lin, Richard Kim, Rajasimman Madhivanan, Arnie Sen, and Dinesh Manocha. Loc-zson: Language-driven object-centric

zero-shot object retrieval and navigation. In *International Conference on Robotics and Automation (ICRA)*, 2024. 1

[23] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. 4, 6

[24] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. 3

[25] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, 2023. 8

[26] Stephen C Levinson. Frames of reference and molyneux's question: Crosslinguistic evidence. 1996. 2

[27] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 5, 6, A1

[28] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 5, 6

[29] 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

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 3, 4, 5, A1

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[32] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 3

[33] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[34] Mustafa Ozuysal, Vincent Lepetit, and Pascal Fua. Pose estimation for category specific multiview object localization. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 778–785. IEEE, 2009. 3

[35] Jiahuan Pei, Irene Viola, Haochen Huang, Junxiao Wang, Moonisa Ahsan, Fanghua Ye, Jiang Yiming, Yao Sai, Di Wang, Zhumin Chen, et al. Autonomous workflow for multimodal fine-grained training assistants towards mixed reality. *arXiv preprint arXiv:2405.13034*, 2024. 1

[36] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 5, 6, A1

[37] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[38] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*. 3

[39] 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

[40] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 1

[41] G Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024). *URL https://goo. gle/GeminiV1-5*. 3, 5

[42] 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, 3, 6

[43] Giuseppe Vallar, Elie Lobel, Gaspare Galati, Alain Berthoz, Luigi Pizzamiglio, and Denis Le Bihan. A fronto-parietal system for computing the egocentric spatial frame of reference in humans. *Experimental brain research*, 124:281–286, 1999. 1

[44] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[45] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281, 2023. 1

[46] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15077–15087, 2024. 1

[47] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian,

et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 5, 6, A1, A2

[48] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024. 3

[49] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 5, A2

[50] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 5, 6

[51] Tino Zaehle, Kirsten Jordan, Torsten Wüstenberg, Jürgen Baudewig, Peter Dechent, and Fred W Mast. The neural basis of the egocentric and allocentric spatial frame of reference. *Brain research*, 1137:92–103, 2007. 1

[52] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 3