Appendix

A. Data Collection

A.1. Instruction Data Generation

We generate three types of data using the inference capabilities of the LLaVA1.5 Vicuna 13B model [30]. The model is prompted to produce questions and answers based on a given context, and the outputs are parsed to construct instruction learning data. The prompts corresponding to each data type are described in Table 11.

A.2. Data Statistics

Training Dataset. We collect data from the ImageNet dataset [10]. We then remove ambiguous images for determining orientations, resulting in a dataset of 2,845 images. To ensure a uniform number of evaluation data per orientation class, we separate benchmark data from the source dataset by extracting 50 images per orientation. The remaining 2,445 images are used to construct the training dataset. For each image, we generate three questions corresponding to the three types of data described earlier. As a result, the training dataset consists of a total of 2,445 images, with three instruction data points generated per image, leading to a total of 7,335 data points. The statistics of the training data is shown in Table 5.

Class	**Count(IN)**
Front-Left	1,284
Front	618
Front-Right	1,326
Right	1,689
Back-Right	225
Back	84
Back-Left	270
Left	1,839

Table 5. The number of data samples for each orientation class.

Benchmark Data. Table 6 presents the statistics of orientation classes for the images in our benchmark. The ImageNet [10] and OmniObject3D [47] datasets are controlled to exhibit a uniform distribution, while the remaining datasets show an imbalanced distribution. Each image is used once for the *Choose* and *Freeform* tasks and twice for the *Verify* task, resulting in the distribution of data samples for each task, as shown in Table 7.

A.3. Task Data Details

- *Choose*: We prompt the model to select the direction an object in the image is facing from among eight directional options. This task is designed to evaluate the basic capability of MLLMs to recognize general directions.
- *Verify*: In this task, there may be a bias towards "yes" or "no" answers, which could result in evaluation metrics failing to accurately represent the actual performance. To address this issue, we design two separate verification tasks for each image: one where the correct answer is "yes" and another where the expected answer is "no." Specifically, for a given image, if the orientation of the image is labeled as "right," we first construct a question asking whether the subject is facing "right." Then, we randomly select one of the remaining orientations (excluding "right" from the eight possible orientations) and create a question asking whether the subject is facing the selected orientation.
- *Freeform*: We recognize that expressions for orientation can vary significantly. For example, phrases like "toward the front," "facing forward," or "looking ahead" may all describe the same orientation. To account for these variations, we allow free-form responses and include a "freeform" metric, verified using the GPT-4o API (`gpt-4o-2024-08-06`), as an additional evaluation measure.

The data format for each task is summarized in Table 8.

Class	Images				
	IN	D_3	DN	PA	3D
Front-Left	50	14	197	171	500
Front	50	38	422	517	500
Front-Right	50	29	213	156	500
Right	50	29	367	586	500
Back-Right	50	4	14	10	500
Back	50	2	18	3	500
Back-Left	50	4	17	5	500
Left	50	32	477	640	500

Table 6. Class-wise data statistics of collected images in our benchmark. Our benchmark is constructed with manually collected orientation annotations from five datasets, ImageNet (IN) [10], D_3 [19], DomainNet (DN) [36], PACS (PA) [27], and OmniObject3D (3D) [47].

Task	Datasets				
	IN	D_3	DN	PA	3D
Choose	400	152	1,725	2,088	4,000
Verify	800	304	3,450	4,176	8,000
Freeform	400	152	1,725	2,088	4,000

Table 7. Task-wise data statistics of our benchmark. Our benchmark is constructed with manually collected orientation annotations from five datasets, ImageNet (IN) [10], D_3 [19], DomainNet (DN) [36], PACS (PA) [27], and OmniObject3D (3D) [47].

Class	Example
Choose	From the perspective of the camera, which orientation is the {Object} in the photo facing? A. front B. front right C. right D. back right E. back F. back left G. left H. front left. Answer with the option's letter and word from the given choices directly.
Verify	Is the {Object} facing "front right" from the camera's perspective? Answer with "yes" or "no" only.
Freeform	From the perspective of the camera, Answer what orientation the {Object} in the picture is facing.

Table 8. The data format for each task.

A.4. 3D Rendered Images

We utilize the OmniObject3D dataset [47], which includes high-quality, real-scanned 3D objects designed to advance 3D perception, reconstruction, and generation tasks in real-world scenarios. We use a total of 500 3D scans distributed across 11 object categories. Each scan is oriented along a specific axis $(+x, -x, +y, -y, +z, -z)$, and we manually unify the coordinate system to standardize their orientations. Subsequently, we place eight cameras (Front-Left, Front, Front-Right, Left, Right, Back-Left, Back, Back-Right) around the object at 45° intervals with respect to the object's center. For the front and back views, the cameras are tilted downward by approximately 20° to better capture the object's orientation. Through this process, we render a total of 4,000 RGB images with a 448×448 image resolution.

B. Experimental Details

We use GPT-4o to evaluate the models on the *Freeform* task and spatial reasoning. The prompts used for GPT-4o are detailed in Table 9.

C. Further Analysis

C.1. Confusion Matrix

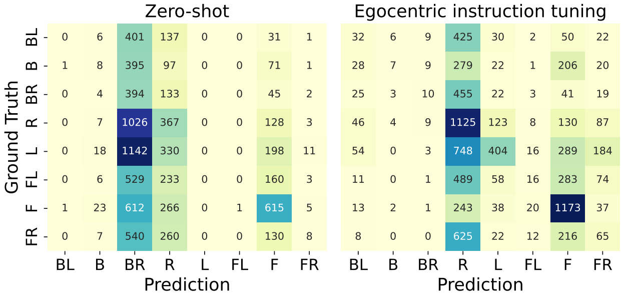
We draw confusion matrix for InternVL [7] in Figure 8. As with the confusion matrices of other models, it can be observed that the responses of the model become more aligned after training, and the biases are mitigated.

C.2. Ablation Test

Table 10 shows the additional results of ablation tests for mPLUG-Owl2 [49] and InternVL [7]. Although not all data types show the same upward trend across every metric for all models, one consistent observation is that the highest performance is achieved when all three response types are utilized. In all tested models, removing any response type resulted in a decrease in task accuracy, confirming that each type contributes to improving the MLLM's understanding of object orientation.

Class	Example
GPT Evaluation	You are given an answer and a prediction representing an object's orientation out of 8 possible directions. Respond with 'yes' if the answer and prediction match, or 'no' if they do not. [Example] If the answer is 'front right' and the prediction is 'facing right while facing the camera,' respond with 'yes.' If the answer is 'front right' and the prediction is 'facing the camera,' respond with 'no,' because 'front' and 'front right' differ in orientation. Answer: {answer} Prediction:{prediction}
Preposition Prompt	From the perspective of the camera, look at the given photo and choose the sentence that best describes its content between the two options. A. {option a} B. {option b}
Preposition GPT Eval	Check if the given prediction matches the ground truth. Respond with 'yes' only if they match, and 'no' otherwise. Letter Answer: A. Sentence Answer: {answer}, Prediction: {prediction}

Table 9. Table with detailed row descriptions



	BL	B	BR	R	L	FL	F	FR
BL	0	6	401	137	0	0	31	1
B	1	8	395	97	0	0	71	1
BR	0	4	394	133	0	0	45	2
R	0	7	1026	367	0	0	128	3
L	0	18	1142	330	0	0	198	11
FL	0	6	529	233	0	0	160	3
F	1	23	612	266	0	1	615	5
FR	0	7	540	260	0	0	130	8

	BL	B	BR	R	L	FL	F	FR
BL	32	6	9	425	30	2	50	22
B	28	7	9	279	22	1	206	20
BR	25	3	10	455	22	3	41	19
R	46	4	9	1125	123	8	130	87
L	54	0	3	748	404	16	289	184
FL	11	0	1	489	58	16	283	74
F	13	2	1	243	38	20	1173	37
FR	8	0	0	625	22	12	216	65

Figure 8. Confusion matrix for the *Choose* task with InternVL.

Response Types			Choose	Verify	Freeform
Type 1	Type 2	Type 3			
✓	✗	✗	25.5	51.7	30.7
✗	✓	✗	18.8	58.3	28.6
✗	✗	✓	21.2	58.9	29.1
✓	✓	✗	23.9	59.3	33.8
✓	✗	✓	28.4	59.9	33.4
✗	✓	✓	21.0	58.1	34.2
✓	✓	✓	28.5	61.5	37.1

Response Types			Choose	Verify	Freeform
Type 1	Type 2	Type 3			
✓	✗	✗	17.4	56.8	18.9
✗	✓	✗	17.9	54.2	32.7
✗	✗	✓	20.7	58.1	41.6
✓	✓	✗	22.1	58.2	31.1
✓	✗	✓	26.5	61.2	43.6
✗	✓	✓	24.3	56.1	44.1
✓	✓	✓	31.4	61.4	48.2

Table 10. Ablation test results for both mPLUG-Owl2 (top) and internVL (bottom). Each response type contributes to performance improvements across all tasks.

Data Type1 Prompt	As a competent assistant, your role is to explain the subparts of the main object and, based on your findings, determine its orientation. These parts of the object should eventually be able to imply some orientation of the object, but questions should never directly include information about its orientation. You must use the information in [Context], but the important thing is that you must find subparts or sub-features of the object given in [Context]. If you identify the main object as a human, you need to find different subparts depending on the orientation the person is facing in the picture. For example, if the person is facing forward (i.e., looking at the camera), you will see the face, eyes, chest, or abdomen. If the person is facing backward, you will see the hip, back, or hair. If the person is facing to the right, you will see one ear and one arm, and you must note that the nose is pointing to the right. For example, [Question]: From camera perspective, does the {sport car} is {facing camera} or {facing away} the camera/observer? 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): headlight, windshield, bumper (Answer the orientation): According to (What I find), The {sport car} facing the camera/observe. [Question]: Does the {a girl} is facing left or facing right from camera perspective? First describe what you can find from the object in the image, then based on that, answer the orientation of the object. [Answer]: (finding): hair, hip, arm, half of nose (Answer the orientation): According to (What I find), The {a girl} facing the left. Now your turn: + context + “[Question]: ” + “[Answer]: ”
Data Type2 Prompt	As a capable vision-language model assistant, your task is to closely examine key features of the object in the provided image and perform the following actions: 1) identify and ask questions about the features that indicate the object’s orientation, and 2) answer the question yourself. To elaborate, you should carefully examine the details that suggest the front or rear of the object, such as eyes, nose, mouth, or tail, or tail lights. Additionally, you should closely check for features that imply a orientation, such as the orientation of the nose, whether one or both arms of a person are visible, or whether the wheels of a car appear as perfect circles, indicating left or right. Instead of making overly general statements, you must create responses that are detailed enough to determine a single orientation. The answer should not just state that something is visible, but rather explain how these features suggest a particular orientation and provide specific details to justify the object’s orientation. For example, you could write something like this: [Question]: Describe in detail the features in the image that indicate the orientation of the object. [Answer]: The two wheels of the car appear perfectly circular, and the car door is visible. This suggests that the object’s orientation is either “left” or “right.” However, since the headlights are on the right side of the image and the red taillights are on the left, the car’s orientation is definitively to the “right.” [Question]: Describe in detail the features in the image that indicate the orientation of the object. [Answer]: Both eyes of the person’s face are visible, but one eye appears larger, and only one cheek is primarily visible, indicating a slanted orientation. The pointed part of the nose is closer to the left side of the image, and the left cheek is not clearly visible from the camera’s perspective. Therefore, the person is facing “front left” Now your turn: + context + “[Question]: ” + “[Answer]: ”
Data Type3 Prompt	As a competent helper like Turn-by-turn navigation, you have the role of understanding the properties of the central object and creating common sense questions and corresponding answer appropriate for them. You must use information in [Context] If you make a question and answer, think carefully that matching the object’s property and common sense with the probable action or capable happening. But pay attention that while making question, you MUST not contain the orientation information directly you get from [Context] in question. Alternatively you can represent with indirect word like corner of image, behind of it, away from camera’s view point or something else. Also, use natural expressions for orientation expressions like facing away from the camera, facing right else. You must now generate a question, similar to the given examples, asking which orientation the object should be turned to face or turn away from the camera with the least angle of rotation, along with the corresponding answer. I’ll give you some example that you can reference. For example, If you find the main object as a sport car, you can make a question like this: [Question]: To make an object face the camera directly with the smallest rotation angle, in which direction should it turn? Choose from [clockwise, counterclockwise, flip, leave as is] and explain why. [Answer]: The sport car is facing to the back, so you have to flip it. [Question]: To make an object face the camera directly with the smallest rotation angle, in which direction should it turn? Choose from [clockwise, counterclockwise, flip, leave as is] and explain why. [Answer]: The sport car is facing to the right, so you have to turn to the counterclockwise to look straight at the camera. Now your turn: + context + “[Question]: ” + “[Answer]: ”

Table 11. Examples of training data for different data types.