AlloEgo-VLM: Disambiguating Allocentric and Egocentric Reference Frames in Vision–Language Models

Kuan-Lin Chen, Yu-Chee Tseng, and Jen-Jee Chen

***Abstract*— This study investigates the challenges of ambiguity faced by Visual–Language Models (VLMs) in understanding spatial semantics. Spatial cognition, influenced by cognitive psychology, spatial science, and cultural contexts, often assigns directionality to objects. In natural language, spatial relationship descriptions frequently omit explicit reference frame specifications, leading to semantic ambiguity. Such ambiguities can lead to erroneous decisions for embodied AI robots. Existing VLMs, due to insufficient annotation of reference frames and object orientations in training data, often produce inconsistent responses. This study derives a new AlloEgo-View dataset consisting of images together with view-specific description of objects in the image with both allocentric and egocentric perspectives. We propose a structured spatial representation method for identifying and annotating key spatial elements in images, including scene descriptions, reference objects and their orientations, target objects and their orientations, as well as reference frame types. Based on AlloEgo-View, we then propose AlloEgo-VLM, a framework that designed a multi-stage training procedure to incorporate these spatial elements into the VLM. Experimental results demonstrate that our approach significantly outperforms state-of-the-art models in spatial orientation reasoning tasks, effectively enhancing the ability of VLMs to resolve spatial semantic ambiguities.**

Keywords: Visual-Language Models, Spatial Semantic Ambiguity, Reference Frame, Egocentric/Allocentric, Multimodal Reasoning

Project page: https://github.com/CKL9001/AlloEgo-VLM

I. Introduction

Understanding spatial semantics is a fundamental capability for Visual-Language Model (VLM), particularly in human-robot interaction and embodied AI for navigation and manipulation tasks [1], [2]. However, spatial descriptions in natural language are inherently ambiguous due to the diversity of reference frames humans may employ. Cognitive psychology and spatial science have shown that humans often omit explicit reference frames in conversation, but instead rely on contexts or cultural conventions [3], [4], [5]. In the example in Fig. 1, the statement *"The person is to the right of the red car"* may be interpreted differently depending on whether an egocentric or allocentric perspective is assumed [6]. Such an ambiguity may persist in general VLMs, if their training data lack explicit spatial reference frame grounding.

Similar challenges have been observed in past works [7] on robot navigation [8], [9], *abstract perspective change* [10], and video action prediction [11]. These studies indeed introduced datasets or benchmarks with annotations of reference frames, providing valuable guidance for spatial

The authors are with the College of Artificial Intelligence, National Yang Ming Chiao Tung University (NYCU), TAIWAN.

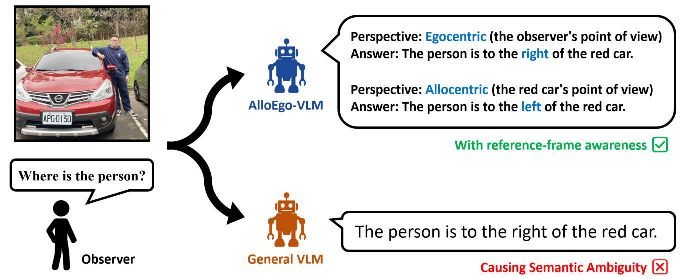


Fig. 1: Orientation ambiguity: egocentric vs. allocentric views.

reasoning. However, their problem formulations often require to explicitly specify the reference frame, either directly or implicitly, in user queries (e.g., *"From the perspective of the observer, where is the person relative to the red car?"*). In other cases, the evaluation is presented as multiple-choice tasks, which do not fully reflect real-world natural language settings. In everyday communications, humans rarely phrase their questions in such a verbose manner.

Our work does not aim to alter the inherent ambiguity that has existed in human language for centuries. Instead, our goal is to enable robots and VLMs to handle such naturally ambiguous queries without requiring users to describe the reference frame explicitly. Even for a simple question like *"Where is the person?"*, the system should automatically filter out inconsistent interpretations and produce the correct answer across different reference frames.

Building on these observations, we note that existing VLMs and embodied AI systems are still rarely trained with explicit annotations of reference frames or object orientations, often leading to inconsistent or even contradictory spatial predictions [12], [13]. Quantitative analyses confirm the severity of this issue: in VSI-Bench (tiny), 71% of errors are due to spatial reasoning, with nearly 22% from egocentric–allocentric transformation [8]; in COMFORT++ and 3DSRBench, baseline VLMs perform close to random under allocentric views (LLaVA-NeXT 48%, SpatialVLM 46%, Molmo 36%), reflecting ambiguity rates of 45–60% [10]. Our own experiments further show that even state-of-the-art models collapse into ambiguity when tested with question-only inputs, yielding reference-frame ambiguity rates above 94% (Table II).

As a result, without being aware of the naturally ambiguous queries within human languages, an embodied robot may easily misinterpreting spatial relations, leading to serious navigation errors, unsafe manipulations, or outright task failures. Although prior works have taken steps toward incorporating spatial reasoning through benchmarks or structured evaluation, the fundamental challenge of resolving **orientation** and **direction ambiguity** in real-world, unconstrained language remains largely unsolved.

To tackle these issues, we take a two-part approach. First, we construct the AlloEgo-View dataset in the form of image–question–answer triplets, making spatial relations and reference frames fully explicit. Second, we introduce a multi-stage training framework that progressively strengthens the model's ability to resolve spatial ambiguity and reason consistently across perspectives. At inference time, the system can interpret naturally ambiguous queries without requiring users to specify the reference frame. The resulting models are lightweight enough for deployment on edge devices such as robots with limited VRAM, ensuring practical applicability in real-world scenarios.

In this work, we make the following contributions:

1) **Revealing spatial direction ambiguity in existing VLMs** – We show that current visual-language models often produce inconsistent spatial reasoning outputs because human descriptions frequently omit explicit reference frames, leaving the models under-constrained during training.
2) **AlloEgo-View: A structured spatial dataset with explicit annotations** – We introduce a dataset and collection methodology that clearly identifies key spatial elements in images, including scene-level descriptions, reference objects and their orientations, target objects and their orientations, and reference frame types with standard relative positions. Each instance provides a fully specified ground truth, ensuring unambiguous supervision and avoiding interpretation gaps.
3) **A DPO-inspired [14] multi-stage iterative training framework** – We propose a fine-tuning pipeline that first aligns the model to the structured spatial representation and then iteratively refines its reasoning ability. Experiments show that this approach, even with only 4k carefully designed samples, outperforms existing state-of-the-art models in spatial orientation reasoning, demonstrating both data efficiency and strong reasoning capability.

II. Related Work

A. Large Language Models and Vision-Language Models

Large language models (LLMs) have shown remarkable progress in tasks such as text summarization, question answering, and multi-step reasoning [19], [20]. Instruction-tuning has further aligned these models with human preferences, improving their robustness and usability in real-world applications [21], [22]. Building on these advances, recent research has extended LLMs into the multimodal domain, giving rise to vision-language models (VLMs).

VLMs such as GPT-4o [23], LLaVA [24], and InstructBLIP [25] integrate visual encoders with powerful language backbones, enabling them to interpret images and answer questions about visual content. These models have demonstrated strong performance in tasks including visual question answering, captioning, and commonsense reasoning over scenes. Their ability to combine linguistic reasoning with visual grounding makes them promising candidates for applications that require understanding spatial relations.

B. Broad Spatial Understanding in Vision-Language Models

Beyond general vision-language reasoning, recent research has explored a wide spectrum of spatial understanding tasks that extend across both 2D and 3D domains. These include predicting object size and relative scale [17], estimating distances and depth relations [26], and reasoning about directions or object localization within complex visual scenes [27]. In 3D settings, spatial reasoning further encompasses navigation-related tasks such as path planning, scene reconstruction, and embodied interaction with objects [7], [28].

To support these tasks, a number of benchmarks and datasets have been developed. CLEVR [15] provides synthetic images with compositional object arrangements and detailed spatial relationships, enabling controlled reasoning tests. GQA [16] extends this to real-world images with structured scene graphs, supporting multi-step reasoning. SpatialSense [17] focuses on natural images with annotated spatial relationships, while SPAR [18] captures relational ambiguity by providing multiple valid interpretations of spatial descriptions. More recent efforts such as ViewSpatial-Bench [9] and Spatial-Comfort [29] further highlight the challenges of reasoning consistently under viewpoint variation.

Despite these advances, most existing approaches still treat spatial reasoning primarily as a geometric or relational problem, without addressing how linguistic descriptions vary depending on perspective or context. This gap motivates a closer examination of allocentric and egocentric perspectives.

C. Allocentric and Egocentric Perspectives in Spatial Reasoning

A central challenge in spatial reasoning lies in distinguishing between two common perspectives: **Egocentric**, which describes spatial relations from the observer's viewpoint, and **Allocentric**, which encodes relations from the perspective of objects themselves, independent of any observer.

Cognitive studies suggest that humans naturally switch between these perspectives depending on context, task demands, and communicative efficiency [30], [31]. However, current VLMs often struggle to robustly handle such variation. Recent datasets such as ViewSpatial-Bench [9] and Spatial-Comfort [29] also highlight the importance of perspective shifts, but their designs remain limited. For instance, they rely on explicit perspective markers in the input (e.g., "from the viewer's perspective ...") or adopt multiple-choice formats—both of which deviate from natural language usage, where speakers typically convey spatial relations without verbose reference-frame indicators.

As a result, VLMs frequently produce inconsistent or contradictory outputs when confronted with short, naturally ambiguous queries. This limitation underscores the need

TABLE I: Comparison of spatial understanding datasets. (✓= yes, $\times = no, \triangle = partial$)

Dataset	Task Description	View-specific	Multi-choice	Comprehensive	Scale
CLEVR [15]	Visual reasoning (synthetic QA)	×	△	×	∼700K
GQA [16]	VQA + scene graph reasoning	×	△	×	∼22M
SpatialSense [17]	Pairwise spatial relation classification	×	×	×	∼11K
SPAR [18]	Spatial perception and reasoning	×	△	×	∼7M
Thinking in Space [8]	Video reasoning and memory	✓	✓	✓	∼5K
Perspective Aware [10]	Abstract perspective change	✓	×	✓	-
SPHERE [7]	Spatial perception and hierarchical reasoning	✓	△	✓	∼2K
ViewSpatial-Bench [9]	Cross-view spatial reasoning	✓	×	✓	∼5K
VLMD4 [11]	Video spatial reasoning	✓	✓	×	∼1K
AlloEgo-View (ours)	Reference frame reasoning	✓	×	×	∼4K

for models that can automatically disambiguate perspective-dependent descriptions and provide stable interpretations across both egocentric and allocentric frames—without requiring explicit cues from users.

A comparison of representative datasets is shown in Table I. "Multi-choice" indicates that the question provides multiple answer options. "Comprehensive" indicates that the question includes sufficient detail (e.g., "If I am standing by the stove and facing the dishwasher, is the refrigerator to my front-left, front-right, back-left, or back-right?").

III. AlloEgo-View: A New Dataset for View-specific Spatial Reasoning

Spatial reference frame, or viewpoint, refers to the coordinate system used to specify the position or orientation of an object. We curate AlloEgo-View, a dataset containing 0.5K manually calibrated and 3.5K automatically generated samples, which together form the training set. Additionally, a separate test set of 1K manually annotated samples is provided. Each item is a view-specific triplet (image, question, answer) designed to mitigate ambiguity between allocentric and egocentric viewpoints. Fig. 2 outlines our methodology, which consists of three steps.

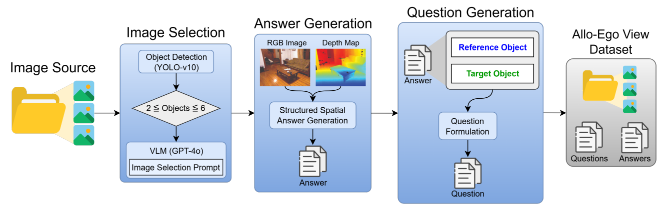


Fig. 2: AlloEgo-View Dataset Pipeline. We construct triplets Image, Question, and Answer within a unified workflow to teach models how to disambiguate reference-frame ambiguity in spatial semantics.

A. RGB Image Selection

We draw images from multiple sources, including GQA [16], SPAR [18], COCO [32], and the NYU Depth Dataset V2 [33]. These images undergo a two-step screening process to guarantee that reference frames are properly represented.

Step 1. Each RGB image is processed with YOLOv10 [34] for object detection. (This step can be skipped if the dataset already includes annotated objects.) An image is retained only if it contains between two and six distinct objects.

Step 2. Each image is evaluated using GPT-4o [35], a vision-language model. A screening prompt asks the VLM to verify whether the image contains between two and six distinct objects and whether the background is relatively clean and uncluttered. The response is binary, and only images that pass this test are retained. Details of the prompt are provided in App. A.

B. Structured Spatial Answer Generation

Next, we generate structured answers from the retained images. The workflow is illustrated in Fig. 3a. We first construct an answer framework and then refine it by adding more view-specific details.

Step 1: Preliminary answer generation. Each above retained image is processed with GPT-4o [35] via a prompt (App. B) to generate an initial answer structured in the following format:

a) Overall image description
b) Reference object <Ref Obj> and Target object <Tgt Obj>
c) Reference object absolute direction <Ref Abs Dir> and Target object absolute direction <Tgt Abs Dir>
d) Egocentric view description
e) Allocentric view description

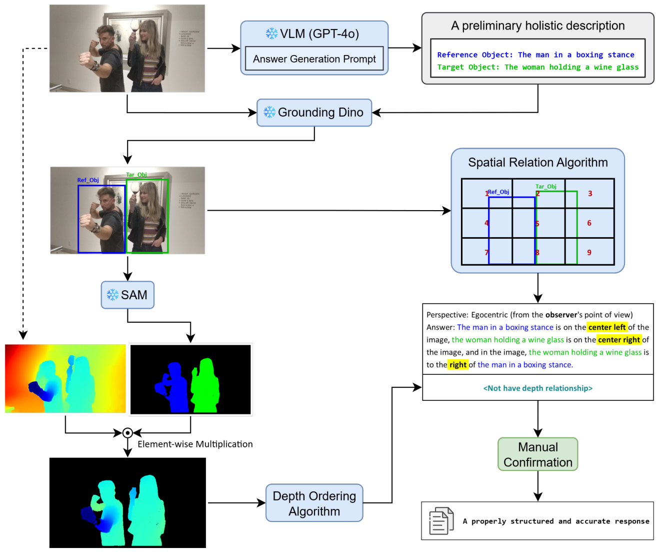
For the detailed format, please refer to Fig. 3b. The egocentric view description includes the statement of

The <Ref Obj> is on the <Ref Pos> of the image.
The <Tgt Obj> is on the <Tgt Pos> of the image.
In the image, <Tgt Obj> is <Ego Rel Dir> of the <Ref Obj>.

The first two statements specify the unique positions of the objects within the image. The subsequent statement, in contrast, describes the relative relationship between the two objects from an egocentric viewpoint. Since the first two statements serve as fixed reference answers, later descriptions from the allocentric viewpoint are not repeated and instead involve only a single specified reference frame:

From the <Ref Obj>'s point of view, the <Tgt Obj> is <Allo Rel Dir> of the <Ref Obj>.

As the VLM may produce erroneous answers, the components of the egocentric view description, namely <Ref Pos>, <Tgt Pos>, and <Ego Rel Dir>, are refined in Step 2. In Step 3, the description is further strengthened by introducing depth statements, thereby capturing the spatial front–back re-



(a)

Overall Image Description: <Overall Image Description>

Reference Object: <Reference Object>
Target Object: <Target Object>

Reference Object Absolute Direction: <Reference Object> is facing **<Direction>**
Target Object Absolute Direction: <Target Object> is facing **<Direction>**

Perspective: Egocentric (from the **observer**'s point of view)
Answer: The <Reference Object> is on the **<Position>** of the image, the <Target Object> is on the **<Position>** of the image, and in the image, <Target Object> is **<Direction>** <Reference Object>.

Perspective: Allocentric (from the <Reference Object>'s point of view)
Answer: From the <Reference Object>'s point of view, the <Target Object> is **<Direction>** of the <Reference Object>.

(b)

Fig. 3: Illustration of our dataset answer design. (a) Demonstrates the process of structured spatial answer generation. (b) Shows the standardized answer format that ensures consistency across egocentric–allocentric tasks.

lationships among objects. Other errors are addressed in Step 4, primarily involving updates to <Ref Abs Dir>, <Tgt Abs Dir>, and the **allocentric view description**. Detailed information regarding this case can be found in App. E (pre-correction, with highlighted segments indicating remaining errors) and App. F (post-correction, showing the accurate results).

Step 2: Egocentric Position and Direction Refinement. To calibrate <Ref Pos>, <Tgt Pos>, and <Ego Rel Dir>, we first prompt Grounding DINO [36] using <Ref Obj>, <Tgt Obj>, and the image as inputs. Grounding DINO helps identify the instance-level grounding of each object, and the outputs are their bounding boxes.

Next, we propose **PosDir Algorithm**. It mainly confirms the absolute and relative positions of the two objects in the image from an egocentric viewpoint:

1) Evenly partition the image into 3×3 grids. Textual descriptions <Ref Pos> and <Ref Pos> are produced for <Ref Obj> and <Tgt Obj> according to the overlapping of their bounding boxes and the grids.
2) Textual description <Ego Rel Dir> is produced according to the relation of the centers of the bounding boxes.

These descriptions are then used to replace the initial answer in Step 1. In the egocentric example in Fig. 3a, three fields (highlighted by yellow) are updated. The complete algorithm is in App. G.

Step 3: Egocentric Depth Ordering Enhancement.

We will enhance the answer by providing more depth information. This is done in three stages. First, we extract the clear shapes of the objects by applying the Segment Anything Model (SAM) [37], using as input the bounding boxes generated in Step 2, to obtain the segmentation maps of <Ref Obj> and <Tgt Obj>. Second, we compute the depth maps of the objects by performing an element-wise multiplication between each object's segmentation map (<Ref Obj> and <Tgt Obj>) and the depth map of the RGB image.

$$Seg(\texttt{<Ref/Tgt Obj>}) \odot Depth(Image)$$

Third, we calculate the mean depth of each object (<Ref Obj> and <Tgt Obj>) based on their depth maps. Comparing the mean depths allows us to determine one of three cases: "<Ref Obj> is deeper," "<Tgt Obj> is deeper," or "no depth relation."

The above result allows us to clearly annotate the depth ordering from the observer's perspective. For example, in Fig. 3a, no depth ordering is present between the man and the woman, so no specific description is generated. Otherwise, an additional statement

> "The <Ref Obj> is between the observer and the <Tgt Obj>." or "The <Tgt Obj> is farther from the observer than the <Ref Obj>."

will be appended to the end of the egocentric viewpoint description. The complete algorithm is in App. H.

Step 4: Manual calibration. In this step, remaining errors are addressed, primarily involving updates to <Ref Abs Dir> and <Tgt Abs Dir>. A manual review is performed, resulting in 0.5K manually corrected answers. We note that <Tgt Abs Dir> in allocentric views is particularly prone to errors, as it heavily depends on the facing direction of <Ref Obj>, which necessitates additional manual effort.The format in Fig. 3b is carefully designed; for some details, please refer to App. I.

C. Structured Spatial Question Generation

To enable AlloEgo-VLM to respond naturally during inference, we associate each above answer with a structured spatial question. We define a set of 17 question templates to capture a wide variety of user queries, ranging from direct relational questions to more conversational descriptions. This ensures linguistic diversity while maintaining semantic consistency. Given any pair (<Ref Obj>, <Tgt Obj>), we show two representative question examples:

a) What is the relationship between the <Tgt Obj> and the <Ref Obj>?
b) Could you describe where the <Tgt Obj> is located?

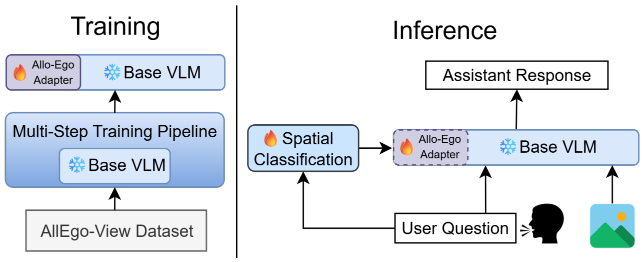


Fig. 4: Overall Methodological Architecture

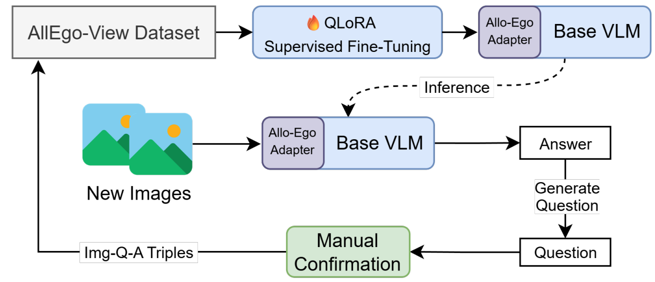


Fig. 5: Multi-stage training pipeline for AlloEgo-Adapter and data expansion of AlloEgo-View.

During dataset construction, one random template is selected for each question. The complete list of 17 templates is provided in APP. D.

IV. AlloEgo-VLM: A Generic VLM with Reference Frame Disambiguation

Our objective is to derive **AlloEgo-VLM** that is able to answer not only questions with reference frame ambiguity, but also general, non-spatial reasoning questions through multi-round natural conversations. To this end, we propose the framework in Fig. 4.

The first step is to fine-tune a base VLM with a **AlloEgo-Adapter** that follows our structured spatial reasoning format. The initial AlloEgo-View dataset, though small in scale, allows us to employ techniques such as quantization and LoRA-based [38] supervised fine-tuning (SFT) and bootstrapped data generalization to gradually expand the scale of AlloEgo-View and train a robust AlloEgo-Adapter. The second step is to train a lightweight **Spatial Classifier** that predicts whether a user query involves reference frame ambiguity. Based on this prediction, the model dynamically decides whether the AlloEgo-Adapter should be engaged. If the query does not involve spatial disambiguation, the original Base VLM directly answers the question. This mechanism enables the final model to flexibly handle both spatial disambiguation and general reasoning tasks, thereby achieving strong generalization performance.

This framework ensures that AlloEgo-VLM can be trained under limited computational resources. In our experiments, training was conducted on a single NVIDIA RTX 3090 GPU, and the tested base VLMs included *Qwen 2.5-VL-7B* [39], *Llama 3.2-V-11B* [40], and *Gemma 3-V-4B* [41].

A. Multi-stage Training of AlloEgo-Adapter

From a frozen base VLM, we train an AlloEgo-Adapter in an iterative manner. During the multi-stage training, we also expand the initial AlloEgo-View dataset from a scale of 0.5K to 4K, inspired by DPO [14]. The pipeline is shown in Fig. 5.

- **Step 1: AlloEgo-Adapter fine-tuning.** We apply QLoRA-based [38] supervised fine-tuning (SFT) on the base VLM using the AlloEgo-View dataset, resulting in AlloEgo-Adapter, a base VLM adapter that aligns with our structured spatial reasoning format.
- **Step 2: Bootstrapped data generation.** Following the procedure in Sec. III-A, we select more images and feed them to the initial model with AlloEgo-Adapter trained in Step 1. Since Step 1 employs full-format SFT, the output structure is strongly constrained, enabling the model to directly produce correct answers from the images even without explicit input questions. Subsequently, using the procedure described in Sec. III-C, we generate their corresponding questions, thereby obtaining more (image, question, answer) triplets conforming to the format of AlloEgo-View.
- **Step 3: Manual confirmation and iterative refinement.** The triplets generated in Step 2 further go through manual confirmation to ensure correctness and natural language expression. The calibrated triplets are then included to expand our AlloEgo-View dataset. Subsequently, we iteratively repeat Steps 1 to 3, until the model converges.

In summary, this pipeline emphasizes two key aspects: (i) progressively expanding AlloEgo-View through bootstrapped generation and manual refinement to a scale of 4K, and (ii) iteratively fine-tuning AlloEgo-Adapter for more robust spatial reasoning. In our setting, we perform five iterations to achieve stable performance.

B. Spatial Classifier

Due to the rigid format of SFT, the adapter-enhanced VLM tends to respond primarily to spatial disambiguation questions. To effectively distinguish view-specific (egocentric/allocentric) from non-view-specific queries, we train a lightweight classifier following the pipeline in Fig. 6.

First, the question is tokenized using a frozen DistilBERT tokenizer [42]. These question tokens are then passed to a frozen DistilBERT encoder [42], producing a question latent representation that preserves the generalizability of the original DistilBERT. Then a classification head consisting of two linear layers follows to make a binary decision of whether the question pertains to spatial-direction reasoning.

DistilBERT [42] was chosen for its efficiency, compact size, and robust semantic encoding capability, making it well-suited for our classification task without requiring full fine-tuning.

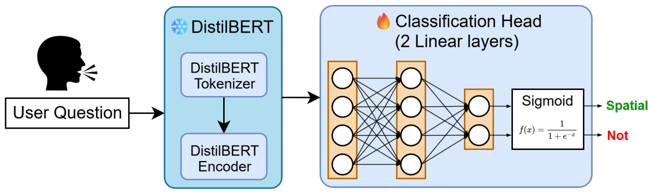


Fig. 6: The spatial classifier, which determines if a question contains reference-frame ambiguity or not.

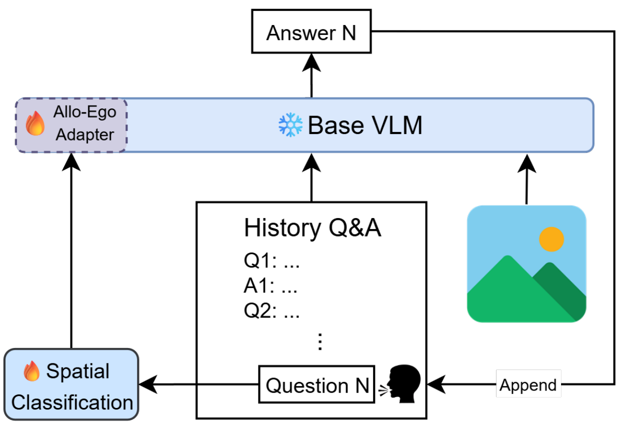


Fig. 7: Inference with Multi-Turn Dialogue. This workflow allows the model to iteratively interact and flexibly switch between egocentric–allocentric disambiguation and general reasoning tasks.

To train the classifier, we curate a dedicated Question-Only dataset, named *AlloEgo-QO*, containing approximately 10K questions without answers. This dataset includes all 4K spatial questions from AlloEgo-View, while the remaining 6K consist of diverse queries collected from online sources (e.g., mathematics, finance, and everyday tasks). The former are treated as positive (i.e., view-related) samples, and the latter as negative samples. We split the dataset into training and test sets using an 8:2 ratio.

C. Support of Multi-round Dialogue

The above modules are integrated to construct the framework shown in Fig. 7, supporting multi-round dialogue. During inference, each new question is first processed by the Spatial Classifier. If the question corresponds to a spatial-direction reasoning task, both the base VLM and AlloEgo-Adapter are invoked; otherwise, only the base VLM is used.

To support multi-round dialogue, the historical conversation log (if any), the query image, and the current question are all input simultaneously to VLM for processing. As mentioned, the invocation of AlloEgo-Adapter depends on the classification result of the Spatial Classifier. The obtained answer, together with the current question, is appended to the historical log to preserve context. Subsequent interactions are processed in a similar iterative manner.

Since the ultimate goal is deployment on a robot, each time the robot observes a new scene, a new dialogue session is initiated without any prior history. Therefore, in our multi-round dialogue setting, the historical log contains only previous questions and answers, while the query image is not re-input for subsequent rounds.

V. Experiments

A. View-Specific Comparisons

In the following, we compare our model deployed on different base VLM models against several state-of-the-art vision-language models, including **GPT-4o** [35], **GPT-4o mini** [35], **Llama 3.2-V** [40], **Gemma 3-V** [41], and variants of **Qwen 2.5-VL** [39]. We consider view awareness and view disambiguation capabilities.

a) Allocentric-Egocentric View Awareness Test: Here, we take the question set of AlloEgo-View derived in Sec. III to perform the experiment. These questions are entirely ambiguous in terms of the questioner's perspective because no reference frame is given in the question. By feeding only these questions to a model without imposing any view-specific format constraint, we intend to test if a model has sufficient view awareness.

Table II reports two metrics. The View Awareness Rate (VAR) measures whether a model's reply explicitly includes a reference frame. Specifically, we follow the evaluation protocol in App. J, where the designed prompt is fed into **GPT-4o** to judge whether the response contains a reference frame. The ambiguity rate measures the proportion of responses in which, when only the **user input (question)** is provided without any explicit reference frame information, the model's reply contains no reference frame and results in inconsistent descriptions. As shown, all evaluated models exhibit extremely high ambiguity rates: **Qwen 2.5-VL (7B)** [39] produces an AR of 94.57%, **Llama 3.2-V (11B)** [40] reaches 100%, and **GPT-4o** [35] variants exceed 99%. These results clearly demonstrate that existing models struggle to resolve spatial relationships based solely on the question text, highlighting the critical importance of incorporating reference frames, structured prompts, or fine-tuned adapters to reduce ambiguity.

The **View-Specific Accuracy (VSA)** measures the models' ability to correctly answer spatial-direction reasoning questions when the reference frame is explicitly provided. Because the sample size is very small, each case can be quickly verified by hand without requiring additional evaluation methods. For most models, this ability remains inconsistent, reinforcing the conclusion that question-only input is insufficient for reliable spatial reasoning. Our SFT-tuned models have strong view awareness. The detailed comparisons are provided in the following discussions.

b) Allocentric-Egocentric View Disambiguation Test: Table III presents a comprehensive performance comparison of our proposed dataset and training methodology against several state-of-the-art (SOTA) vision-language models, including **GPT-4o** [35], **GPT-4o mini** [35], **Llama 3.2-V** [40], **Gemma 3-V** [41], and variants of **Qwen 2.5-VL** [39]. The test dataset of 1K manually annotated samples, as described

TABLE II: View Awareness Rate (VAR). Models are tested with question-only inputs, without restricting response format, to measure the occurrence of reference-frame ambiguity. View-Specific Accuracy (VSA) measures the accuracy of spatial-direction reasoning when the model provides a reference frame.

Method	Params	VAR↓	VSA↑
Qwen 2.5 - VL	7B	94.568%	50%
Llama 3.2 - V	11B	100%	-
Gemma 3 - V	4B	99.753%	100%
GPT - 4o	-	99.259%	66.666%
GPT - 4o mini	-	99.012%	75%

in Sec.III, maintains the same structured format, ensuring consistency and fairness in evaluation.

Three evaluation settings are reported:

1) **View-Specifi Format-only prompt (VS-Format)**: The model receives only the input question along with the expected output format, without additional guidance on spatial semantics or structured reasoning.
2) **View-Specifi Textbook-level prompt (VS-Textbook)**: In addition to the input question, the model receives a detailed prompt (as described in App. B) that guides its reasoning process. This prompt instructs the model on how to identify allocentric versus egocentric descriptions, how to avoid generating ambiguous or contradictory answers, and other aspects of structured spatial reasoning, etc.
3) **Supervised Fine-Tuning (SFT)**: Models are trained using our iterative SFT pipeline described in Sec. IV, which includes structured dataset expansion, Spatial Adapter integration, and multi-turn spatial grounding. This approach ensures that the model consistently produces unambiguous, contextually accurate, and task-aligned outputs.

We report four metrics (as described in the format shown in Fig. 3b): **AD_RO** (Reference Object Absolute Direction), **AD_TO** (Target Object Absolute Direction), **Ego** (egocentric), and **Allo** (allocentric). Each entry in our test dataset is annotated with a **ground-truth answer**. To accommodate the VLM's diverse responses, we use **GPT-4o** [35]] to automatically evaluate the similarity between generated answers and the ground truth, assigning scores from 1 to 10 (see scoring prompt in App. C).

From Table III, several key observations can be drawn:

- Within each model, a clear stepwise improvement is observed: **VS-Format** yields only moderate performance, **VS-Textbook** brings substantial gains across all metrics, and the full **SFT** setting consistently delivers the best results. This progressive trend demonstrates that moving from format-only prompting to structured prompting, and ultimately to fine-tuned training, markedly enhances spatial reasoning ability.
- When comparing the highlighted gray rows, it is notable that open-source models such as **Qwen 2.5-VL (7B)** and **Llama 3.2-V (11B)**, which can be directly fine-tuned via SFT, achieve performance that not only significantly surpasses their own VS-Format and VS-Textbook settings but also exceeds the strongest VS-Textbook baseline, **GPT-4o**. This result highlights that our fine-tuning strategy enables smaller open-source models to outperform larger proprietary SOTA systems.
- The superior performance of our SFT approach can be attributed to several factors: (i) the dataset explicitly encodes reference objects, target objects, and directional relationships, thereby reducing ambiguity; (ii) iterative fine-tuning with human verification ensures that the model acquires consistent spatial reasoning patterns; and (iii) the Spatial Adapter enables the base model to generalize effectively to diverse spatial configurations while maintaining structured output formats.

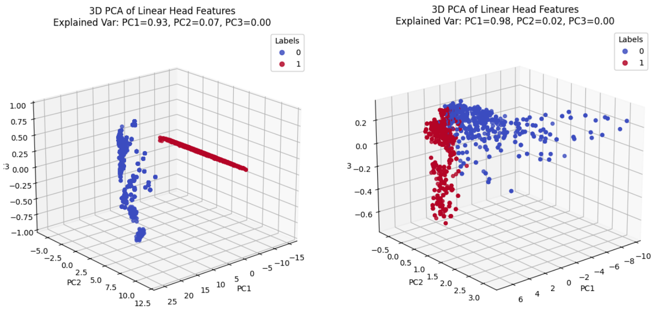
Overall, these results demonstrate that our dataset and training methodology substantially enhance VLMs' ability to perform precise spatial reasoning and generate contextually accurate, unambiguous outputs.

TABLE III: Model Performance Comparison. Performance of **Qwen 2.5-VL**, **Llama 3.2-V**, **Gemma 3-V**, **GPT-4o**, and **GPT-4o-mini** under three prompting/training settings: **VS-Format** (View-Specific Format-only prompting), **VS-Textbook** (View-Specific Textbook-level prompting), and **SFT** (Supervised Fine-Tuning).

Method	Params	AD_RO↑	AD_TO↑	Ego↑	Allo↑
Qwen 2.5-VL (VS-Format)	7B	4.08	3.62	4.51	2.65
Qwen 2.5-VL (VS-Textbook)	7B	5.22	4.61	5.39	4.16
Qwen 2.5-VL (SFT)	7B	**7.94**	**8.04**	**8.19**	**6.25**
Llama 3.2-V (VS-Format)	11B	3.82	3.53	4.36	3.74
Llama 3.2-V (VS-Textbook)	11B	4.95	4.56	5.03	3.38
Llama 3.2-V (SFT)	11B	**7.92**	**8.17**	**8.28**	**6.72**
Gemma 3-V (VS-Format)	4B	3.59	3.20	1.84	3.94
Gemma 3-V (VS-Textbook)	4B	4.75	4.50	4.89	3.86
Gemma 3-V (SFT)	4B	5.38	5.76	4.36	4.30
GPT-4o mini (VS-Format)	-	4.40	4.16	5.67	4.43
GPT-4o mini (VS-Textbook)	-	6.32	6.41	7.02	5.27
GPT-4o (VS-Format)	-	4.54	4.56	5.95	5.05
GPT-4o (VS-Textbook)	-	**7.25**	**7.42**	**7.34**	**5.91**

TABLE IV: Validation results of the spatial classifier against other VLMs, reporting parameter count, model size, inference time, and accuracy.

Method	Para.	Model Size	Inf. Time	Acc.
Spatial Classifier	6M	0.25GB	≈2ms	100%
Qwen 2.5-VL	7B	7GB	≈260ms	99.53%
Llama 3.2-V	11B	11GB	≈700ms	99.71%
Gemma 3-V	4B	4GB	≈240ms	91.24%



(a) Training set

(b) Test set

Fig. 8: 3D PCA projection of features encoded by the spatial classifier.

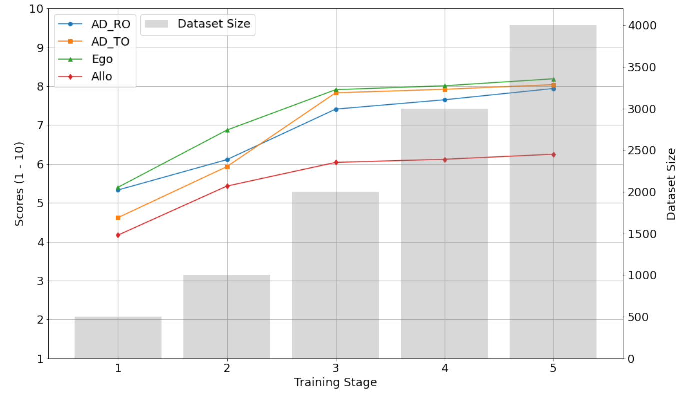


Fig. 9: Performance trajectory of AlloEgo-Adapter on Qwen 2.5-VL 7B during progressive training. The line plot shows performance scores, and the bar chart indicates dataset sizes.

B. *Performance of the Spatial Classifier*

The spatial classifier plays a key role in preventing AlloEgo-VLM from overfitting to spatial questions. We evaluate it on the AlloEgo-QO test set, with results summarized in Table IV, alongside several state-of-the-art VLMs. Our model achieves perfect accuracy (100%) while maintaining a lightweight design (6M parameters, 0.25GB) and extremely fast inference speed (≈2ms). Since this is a relatively simple task, Qwen 2.5-VL and Llama 3.2-V also achieve accuracies above 99%, while Gemma 3-V lags behind on this binary decision task. These results highlight the superior efficiency and reliability of our classifier in distinguishing spatial reasoning questions from others. In contrast, existing VLMs are not optimized for such binary discrimination, as they are designed for general-purpose question answering. Thus, our lightweight spatial classifier effectively serves its purpose in our setting.

To further examine the learned representations, we visualize the output features of the final linear M3 layer using 3D PCA [43]. As shown in Fig. 8, the representations of the two classes are clearly separated in both training and unseen test data, confirming that the classifier has learned highly discriminative latent features.

C. *Training Performance of AlloEgo-Adapter*

AlloEgo-Adapter was trained progressively in a multi-stage manner, with each stage generating additional view-specific data for fine-tuning. Fig. 9 shows the performance trajectory using Qwen 2.5-VL 7B as the backbone. The model improves steadily in the early stages and begins to plateau after Stage 3. The bar chart illustrates the dataset growth during bootstrapped data generation, ranging from 0.5K samples at Stage 1 to 4K samples at Stage 5. At Stage 5, where the training dataset reaches 4K samples, performance stabilizes. These results indicate that AlloEgo-VLM can be effectively trained with a relatively small dataset of 4K samples, achieving performance levels that approach or surpass 8 (out of 10) across the majority of evaluation metrics.

Qwen 2.5-VL 7B was selected empirically for bootstrapped data generation, as it offers a favorable balance between capacity and performance. As shown in Table III, training AlloEgo-Adapter on the curated AlloEgo-View dataset also significantly enhances the view disambiguation capabilities of Llama 3.2-V and Gemma 3-V.

D. *Additional Experiments*

Several other experiments, with illustrations, are provided in the supplementary materials.

a) Questions with Non-existing Objects: We observed an intriguing phenomenon during the early training stages of **AlloEgo-Adapter**. When prompted with references to objects that do not actually exist in the image, the model tended to generate erroneous object descriptions purely based on the visual context, without considering whether the referenced object was present in the query. This behavior results in misleading responses, as users may be unaware that the mentioned object is not in the scene, which in safety-critical applications could potentially lead to dangerous situations (refer to App. K, Fig. 10a).

b) Questions with pure text: Similarly, during the early training stages of **AlloEgo-Adapter**, when the prompt consisted solely of a textual question without any accompanying image input, the model generated hallucinatory descriptions following the output format shown in Fig. 3b (refer to App. K, Fig. 10b). Such behavior can be problematic in practical deployments: if the camera input fails but remains unnoticed, the robot may continue to act upon these fabricated responses, potentially leading to hazardous consequences.

As the training stage progressed to Stage 4 (with training datasets exceeding 3K samples), Qwen2.5-VL-7B [39] exhibited the emergence of correct responses for both scenarios: identifying that an object is absent from the image, and providing valid answers to pure text questions. In contrast, when trained on datasets smaller than 3K samples, the model tended to rigidly follow the prescribed answer format, even producing hallucinated responses. This behavior indicates that insufficient training signals caused the model to rely heavily on format imitation.

Under strong format supervision, one might expect the model to be "locked" into rigid format adherence. However, our results demonstrate that, given sufficient data, the model can autonomously determine when to apply prior knowledge to answer correctly, indicating genuine learning of spatial semantics rather than mere format imitation. Notably, the model was never explicitly trained to handle these two types of questions, nor did such cases appear in the training dataset, further underscoring that the observed behaviors emerge from the model's learned generalization rather than dataset memorization.

c) Structured Response Design: Our structured spatial format is strictly enforced across all responses to effectively prevent view ambiguities. As summarized in App. I, the following principles are applied to minimize inconsistencies and ensure clarity:

1) **Global Context:** Begin with an overall image description to establish a coherent understanding before detailed reasoning.
2) **Absolute Directions:** Provide `<Ref Abs Dir>` and `<Tgt Abs Dir>` to clarify relative directions and improve consistency in both egocentric and allocentric views.
3) **Object Positions:** Specify `<Ref Pos>` and `<Tgt Pos>` to avoid ambiguities when objects appear at distant image boundaries.
4) **Egocentric Grounding:** Encourage phrasing tied explicitly to the observer's perspective (e.g., "In the image, …") to anchor reasoning.
5) **Depth Ordering:** Replace vague front–back terms with explicit relational statements (e.g., "between the observer and …") to reduce 3D ambiguity.
6) **Allocentric Challenges:** Recognize that objects like sofas or monitors lack intrinsic orientation, underscoring the limitations of current VLMs.
7) **Simplifying Labeling:** Avoid labeling cases such as "looking back," where cultural or individual interpretations differ.
8) **Non-directional Objects:** Exclude objects without intrinsic directions (e.g., balls, plants, fruits) from annotation.
9) **Geographic Directions:** Exclude absolute geographic references (e.g., north, south, east, west) as they lie outside the scope of this work.

These techniques highlight the necessity of rigid structural enforcement: without such constraints, variations across individuals and cultural backgrounds can lead to substantial inconsistencies in spatial reasoning.

VI. Conclusions and Future work

In this work, we investigated the challenge of spatial semantic ambiguity in visual-language models (VLMs), focusing on the distinctions between allocentric and egocentric reference frames. We introduced a structured spatial representation framework, a corresponding dataset, and a multi-stage iterative fine-tuning pipeline leveraging QLoRA [38] , which together enable VLMs to accurately interpret spatial relationships and reduce ambiguity.

Experimental results demonstrate that our approach significantly outperforms existing state-of-the-art models, including GPT-4o [35], Llama3.2-Vision [40], and Gemma3-Vision [41], in spatial reasoning tasks. Notably, Qwen2.5-VL-7B [39] trained on the sufficiently large dataset exhibits emergent behavior, such as identifying that an object is absent from the image and providing valid answers to pure text questions, while maintaining robust performance across different query types.

Our findings also highlight the importance of dataset scale and structured supervision: training on an insufficient dataset often causes the model to over-rely on format imitation, whereas a sufficiently large dataset with careful annotations enables the emergence of genuine reasoning capabilities. Furthermore, the Spatial Classification model achieves perfect separation of spatial versus non-spatial questions, and 3D PCA visualizations confirm highly discriminative latent representations.

Overall, this study provides a practical and efficient methodology for enhancing VLMs' spatial reasoning and demonstrates the value of structured spatial supervision in reducing ambiguity, enabling reliable deployment in applications such as robotics, human-robot interaction, and multimodal reasoning.

The training framework and workflow described in Sec. IV can be rapidly deployed across different VLMs. If a VLM encounters difficulties on other tasks, a corresponding Else-Adapter can be trained following the procedure in Sec. IV-A, and the Classifier in Sec. IV-B can be updated to identify different question types. Combined with the Multi-turn Inference mechanism in Sec. IV-C, the system can flexibly switch between Adapters, enabling extended capabilities. Unlike traditional MOE (Mixture of Experts)[44] approaches, which require loading multiple models simultaneously for inference and can create resource bottlenecks on edge devices such as robots, our method loads only a single model at a time while dynamically inserting the relevant Adapter, achieving an efficient balance between performance and resource usage.

References

[1] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 14 455–14 465.

[2] B. Ji, S. Agrawal, Q. Tang, and Y. Wu, "Enhancing spatial reasoning in vision-language models via chain-of-thought prompting and reinforcement learning," *arXiv preprint arXiv:2507.13362*, 2025.

[3] G. Janzen, D. B. M. Haun, and S. C. Levinson, "Tracking down abstract linguistic meaning: Neural correlates of spatial frame of reference ambiguities in language," *PLoS ONE*, vol. 7, no. 2, p. e30657, 2012.

[4] S. C. Levinson, "Reference frames in language and cognition: Cross-population mismatches," in *Reference frames in language and cognition: cross-population mismatches*. De Gruyter, 2022.

[5] A. Majid, M. Bowerman, S. Kita, D. B. Haun, and S. C. Levinson, "Can language restructure cognition? the case for space," *Trends in Cognitive Sciences*, 2004, discusses spatial FoR differences across languages.

[6] F. Filimon, "Are all spatial reference frames egocentric? reinterpreting evidence for allocentric, object-centered, or world-centered reference frames," *Frontiers in Human Neuroscience*, vol. 9, p. 648, 2015.
[7] W. Zhang, W. E. Ng, L. Ma, Y. Wang, J. Zhao, A. Koenecke, B. Li, and L. Wang, "Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation," *arXiv preprint arXiv:2412.12693*, 2024.
[8] J. Yang, S. Yang, A. W. Gupta, F.-F. Li, and S. Xie, "Thinking in space: How multimodal large language models see, remember, and recall spaces," *arXiv preprint arXiv:2412.14171*, 2024.
[9] D. Li, H. Li, Z. Wang, Y. Yan, H. Zhang, S. Chen, G. Hou, S. Jiang, W. Zhang, Y. Shen, W. Lu, and Y. Zhuang, "Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models," *arXiv preprint arXiv:2505.21500*, 2025.
[10] P. Y. Lee, J. Je, C. Park, M. A. Uy, L. Guibas, and M. Sung, "Perspective-aware reasoning in vision-language models via mental imagery simulation," *arXiv preprint arXiv:2504.17207*, 2025.
[11] S. Zhou, A. Vilesov, X. He, Z. Wan, S. Zhang, A. Nagachandra, D. Chang, D. Chen, X. E. Wang, and A. Kadambi, "Vlm4d: Towards spatiotemporal awareness in vision language models," *arXiv preprint arXiv:2508.02095*, 2025.
[12] I. Stogiannidis, S. McDonagh, and S. A. Tsaftaris, "Mind the gap: Benchmarking spatial reasoning in vision-language models," *arXiv preprint arXiv:2503.19707*, 2025.
[13] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, Y. Li, and N. Joshi, "Is a picture worth a thousand words? delving into spatial reasoning for vision language models," *arXiv preprint arXiv:2406.14852*, 2024.
[14] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *CoRR*, vol. abs/2305.18290, 2023.
[15] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[16] D. A. Hudson and C. 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 (CVPR)*, 2019.
[17] K. Yang, O. Russakovsky, and J. Deng, "Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition," in *International Conference on Computer Vision (ICCV)*, 2019.
[18] J. Zhang, Y. Chen, Y. Zhou, Y. Xu, Z. Huang, J. Mei, J. Chen, Y.-J. Yuan, X. Cai, G. Huang, X. Quan, H. Xu, and L. Zhang, "From flatland to space: Teaching vision-language models to perceive and reason in 3d," *arXiv preprint arXiv:2503.22976v1*, 2025.
[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[20] H. Touvron, T. Lavril, G. Izacard, X. Martinet, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
[21] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, *et al.*, "Finetuned language models are zero-shot learners," in *International Conference on Learning Representations (ICLR)*, 2022.
[22] L. Ouyang, J. Wu, X. Jiang, D. Almeida, *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
[23] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[24] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
[25] W. Dai, J. Li, D. Li, A. D. Bagdanov, *et al.*, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
[26] Y.-H. Liao, R. Mahmood, S. Fidler, and D. Acuña, "Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models," *arXiv preprint arXiv:2409.09788*, 2024.
[27] A.-C. Cheng, H. Yin, Y. Fu, *et al.*, "Spatialrgpt: Grounded spatial reasoning in vision language models," *arXiv preprint arXiv:2406.01584*, 2024.
[28] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, S. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2011.
[29] Z. Zhang, F. Hu, J. Lee, F. Shi, P. Kordjamshidi, J. Chai, and Z. Ma, "Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities," *arXiv preprint arXiv:2410.17385*, 2024.
[30] R. Orti, T. Iachini, E. D' Agostino, F. Ruotolo, and G. Ruggiero, "Cognitive load in switching between egocentric and allocentric spatial frames of reference: a pupillometry study," *Scientific Reports*, 2025.
[31] A. Alexander, "Switching between egocentric and allocentric coordinates," Interview (Q&A) at Sainsbury Wellcome Centre Seminar, 2023.
[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.
[33] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 746–760.
[34] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.
[35] A. Hurst, A. Lerer, A. P. Goucher, and et al., "Gpt-4o system card," Tech. Rep., 2024.
[36] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Zhang, H. Su, J. Zhu, X. Du, L. Zhang, and M. Z. Shou, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.
[37] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.
[38] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
[39] Qwen Team, "Qwen2.5-vl technical report," arXiv, CoRR abs/2502.13923, Feb. 2025, introduces the Qwen2.5-VL vision-language model series (3B, 7B, 72B), featuring window attention, dynamic FPS sampling, and enhanced temporal reasoning. The 7B-parameter model balances speed and performance in multimodal tasks.
[40] Meta AI, "Llama 3.2-v: Meta's multimodal vision-language model," *arXiv preprint arXiv:2409.12345*, 2024.
[41] Google DeepMind, "Gemma 3-v: Google's multimodal vision-language model," *arXiv preprint arXiv:2503.19786*, 2025.
[42] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
[43] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.
[44] A. Q. Jiang, A. Sablayrolles, A. Roux, and et al., "Mixtral of experts," 1 2024, arXiv:2401.04088 [cs.LG].

Appendix

A. Prompt for RGB Image Selection

The prompt for RGB image selection in Sec. III-A is shown below:

```
Please determine whether the image meets
the following criteria:
 1) The image contains approximately 2 to
    6 distinct, identifiable objects or
    entities.
 2) The background is relatively clean and
    uncluttered.
 3) The scene could potentially lead to
    referential ambiguity in natural language
    descriptions due to varied perspectives or
    viewpoints.
These images will be used to generate
question-answer pairs related to referential
ambiguity. Respond only with "Yes" or "No".
```

B. Prompt for Generating Structured Initial Answers

We employ an LLM (GPT-4o) with this prompt to generate preliminary Assistant Responses during dataset construction.

Prompt:

Please follow the instructions below to describe the direction and spatial relationship between two objects:

1) **Overall Image Description:** Based on the visual content of the image, provide a comprehensive description of the entire scene, including:
   - The overall setting and background context (e.g., indoor/outdoor, urban/natural environment).
   - The number and types of visible objects (e.g., people, vehicles, buildings, furniture).
   - The spatial distribution and relative positions of objects within the scene (e.g., clustered on the left, evenly spread across the image).
   - Any prominent visual structures or compositional features that help define the spatial layout (e.g., roads, walls, floor lines, depth cues in the background).
2) **Reference Object Absolute Orientation:**
   - Do not compare it to any other object.
   - Only describe its own directional properties.
   - Use absolute terms such as: "facing left", "facing right", "facing upward", "facing downward", "facing the observer", "facing away from the observer", or "this object has no inherent direction".
   - An object is considered to have inherent directionality if its front, back, left, or right side can be visually inferred from its shape, posture, or design (e.g., a person, car, or animal). Objects like chairs or cups may have direction depending on their orientation. If no such direction is visually evident, state "this object has no inherent direction".
3) **Target Object Absolute Orientation:**
   - Same rules as the Reference Object: do not compare to any other object and only describe its own directionality.
4) **Relative Position of the Target Object with Respect to the Reference Object:**
   - If at least one of the two objects has directionality, provide both **Egocentric** and **Allocentric** perspectives.
   - If neither of the two objects has directionality, provide **Egocentric** only.
   - **Egocentric Description (Observer-Centered):**
     a) Treat both the Reference Object and the Target Object as 2D bounding boxes in screen space.
     b) Think of the image as a nine-square grid. Describe their **individual screen positions** using: "upper left", "upper center", "upper right", "center left", "center", "center right", "lower left", "lower center", "lower right".
     c) Describe their **relative screen positions** using:

        "<Target Object> is at the same position as the <Reference Object>", "<Target Object> is above the <Reference Object>", "<Target Object> is below the <Reference Object>", "<Target Object> is to the left of the <Reference Object>", "<Target Object> is to the right of the <Reference Object>", "<Target Object> is to the upper left of the <Reference Object>", "<Target Object> is to the lower left of the <Reference Object>", "<Target Object> is to the upper right of the <Reference Object>", "<Target Object> is to the lower right of the <Reference Object>".
     d) If a depth relationship is visible (occlusion or perspective cues), indicate:

        "<Target Object> is between the observer and the <Reference Object>", "<Reference Object> is closer to the observer than the <Target Object>", or "<Target Object> is farther from the observer than the <Reference Object>".
   - **Allocentric Description (Reference Object-Centered):**
     – Use the orientation of the Reference Object to describe the location of the Target Object.
     – Use spatial terms such as: in front of, behind, to the left of, to the right of, diagonally front-left, etc.
5) **Response Format:**

   Overall Image Description: <Overall Image Description>

   Reference Object: <Reference Object>
   Target Object: <Target Object>

   Reference Object Absolute Direction: <Reference Object> is facing <Direction>
   Target Object Absolute Direction: <Target Object> is facing <Direction>

   Perspective: Egocentric (from the observer's point of view)
   Answer: The <Reference Object> is on the <Position> of the image, the <Target Object> is on the <Position> of the image, and in the image, <Target Object> is <Direction> <Reference Object>.

   Perspective: Allocentric (from the <Reference Object>'s point of view)
   Answer: From the <Reference Object>'s perspective, the <Target Object> is <Direction> of the <Reference Object>.

Example 1:

Overall Image Description: The image shows a heartwarming scene of a golden retriever lying on the floor next to a gray and white cat. The cat is gently nuzzling the dog's face, creating a sense of affection between the two pets. In front of them is a white bowl, likely containing food, suggesting they might be sharing a meal. The setting appears to be a cozy indoor space with white cabinets, shelves with books or papers, and a light-colored floor.
Reference Object: Dog
Target Object: Cat

Reference Object Absolute Direction: The dog is facing the observer
Target Object Absolute Direction: The cat is facing left

Perspective: Egocentric (from the observer's point of view)
Answer: The dog is on the left middle of the image, the cat is on the right middle of the image, and in the image, cat is to the right of dog.

Perspective: Allocentric (from the Dog's point of view)
Answer: From the dog's perspective, the cat is on its right side.

Example 2:

Overall Image Description: This image shows two plastic buckets placed on a smooth, light-colored surface. The bucket in the foreground is blue with a white handle, while the bucket behind it is red, also with a white handle. The blue bucket is positioned slightly to the left and in front of the red one, creating a clear sense of depth and perspective. The scene appears to be well-lit, likely photographed indoors or in a shaded outdoor setting. The overall composition is simple and minimalistic.

Reference Object: Red bucket
Target Object: Blue bucket

Reference Object Absolute Direction: The red bucket has no inherent direction
Target Object Absolute Direction: The blue bucket has no inherent direction

Perspective: Egocentric (from the observer's point of view)
Answer: The red bucket is on the upper right of the image, the blue bucket is on the center left of the image, and in the image, the red bucket is to the upper right of the blue bucket.

Perspective: Allocentric (from the red bucket's point of view)
Answer: No Allocentric.

C. Prompt for Automated Model Scoring

GPT-4o uses this prompt to assess the quality of model outputs, providing quantitative evaluation for Table III.

Prompt:
You are a semantic evaluation expert.

Assistant Response: {AR["answers"][i]}
Ground Truth Answer: {GT["answers"][i]}

Evaluate their similarity in four specific aspects:

1) **Reference Object Absolute Direction Score: X / 10**
   - Assess how accurately the Reference Object Absolute Direction in the Assistant Response matches the Ground Truth Answer.
   - Consider semantic correctness, direction consistency, and clarity.
2) **Target Object Absolute Direction Score: X / 10**
   - Assess how accurately the Target Object Absolute Direction in the Assistant Response matches the Ground Truth Answer.
   - Consider semantic correctness, direction consistency, and clarity.
3) **Egocentric Answer Score: X / 10**
   - Assess how accurately the Egocentric answer in the Assistant Response matches the Ground Truth Answer.
   - Consider whether the spatial relationship and positional details match.
4) **Allocentric Answer Score: X / 10**
   - Assess how accurately the Allocentric answer in the Assistant Response matches the Ground Truth Answer.
   - Consider whether the description from the reference object's perspective is semantically correct and consistent.

For each category, score from 1 to 10 (10 = completely accurate and aligned; 1 = entirely incorrect).

Then, for each category, provide a clear explanation of your reasoning, addressing:

- Whether meanings align
- Whether any important details are missing or incorrect
- Whether the response is misleading
- Clarity of expression

Output Format:
Reference Object Absolute Direction Score: X / 10
Explanation: ...

Target Object Absolute Direction Score: X / 10
Explanation: ...

Egocentric Answer Score: X / 10
Explanation: ...

Allocentric Answer Score: X / 10

Explanation: ...

D. Question Template List

User inputs are generated from 17 predefined templates to provide consistent yet diverse spatial queries.

1) Where is the <Target Object> in relation to the <Reference Object>?
2) How are the <Target Object> and <Reference Object> positioned?
3) Can you describe where the <Target Object> is?
4) How would you describe the position of the <Target Object> compared to the <Reference Object>?
5) What is the location of the <Target Object> relative to the <Reference Object>?
6) Where would you say the <Target Object> is placed?
7) Tell me how the <Target Object> and the <Reference Object> are arranged.
8) If someone asked you where the <Target Object> is, what would you say?
9) Where is the <Target Object> located with respect to the <Reference Object>?
10) What is the spatial relationship between the <Target Object> and the <Reference Object>?
11) Can you point out where the <Target Object> is compared to the <Reference Object>?
12) Where do you see the <Target Object>?
13) What is the position of the <Target Object> in relation to the other object?
14) Where does the <Target Object> appear to be?
15) Which side of the <Reference Object> is the <Target Object> on?
16) How would you explain where the <Target Object> is to someone else?
17) Looking at the scene, where is the <Target Object>?

E. A preliminary holistic description

Generated by GPT-4o [35] via the Answer Generation Prompt: A preliminary holistic description

```
Overall Image Description: The image shows two people indoors, standing in front of a framed artwork on a wall. The man in a boxing stance is on the left, and the woman holding a wine glass is on the right. The setting appears to be a hallway or lobby area, with a sign on the wall listing various locations such as a roof garden and lounge.

Reference Object: The man in a boxing stance
Target Object: The woman holding a wine glass

Reference Object Absolute Direction: The man in a boxing stance is facing the observer
Target Object Absolute Direction: The woman holding a wine glass is facing the observer

Perspective: Egocentric (from the observer's point of view)
Answer: The man in a boxing stance is on the left of the image, the woman holding a wine glass is on the right of the image, and in the image, the woman holding a wine glass is to the right of the man in a boxing stance.

Perspective: Allocentric (from the man in a boxing stance's point of view)
Answer: From the man in a boxing stance's point of view, the woman holding a wine glass is to the right of the man in a boxing stance.
```

F. A properly structured and accurate response

Generated by Structured Spatial Answer Generation: A properly structured and accurate response.

```
Overall Image Description: The image shows two people indoors, standing in front of a framed artwork on a wall. The man in a boxing stance is on the left, and the woman holding a wine glass is on the right. The setting appears to be a hallway or lobby area, with a sign on the wall listing various locations such as a roof garden and lounge.

Reference Object: The man in a boxing stance
Target Object: The woman holding a wine glass

Reference Object Absolute Direction: The man in a boxing stance is facing the observer
Target Object Absolute Direction: The woman holding a wine glass is facing left and facing the observer

Perspective: Egocentric (from the observer's point of view)
Answer: The man in a boxing stance is on the center left of the image, the woman holding a wine glass is on the center right of the image, and in the image, the woman holding a wine glass is to the right of the man in a boxing stance.

Perspective: Allocentric (from the man in a boxing stance's point of view)
Answer: From the man in a boxing stance's point of view, the woman holding a wine glass is to the left of the man in a boxing stance.
```

G. Spatial Relation Algorithm

Algorithm 1 Grid-Based Spatial Relation Extraction

Require: Reference object bounding box $BBOX_r = (x_{r1}, y_{r1}, x_{r2}, y_{r2})$, Target object bounding box $BBOX_t = (x_{t1}, y_{t1}, x_{t2}, y_{t2})$, Grid indices $grids \subseteq \{1, \ldots, 9\}$

Ensure: Description of the reference object d_r, Description of the target object d_t, Relative relation d_{rel}

Determine the grid index set $g_r \subseteq grids$ covered by the rectangle $BBOX_r = (x_{r1}, y_{r1}, x_{r2}, y_{r2})$

Determine the grid index set $g_t \subseteq grids$ covered by the rectangle $BBOX_t = (x_{t1}, y_{t1}, x_{t2}, y_{t2})$

if $g_r = \emptyset$ or $g_t = \emptyset$ **then**

  $d_r \leftarrow$ "Unknown", $d_t \leftarrow$ "Unknown"

else

  Look up the corresponding description for g_r in Table V, assign to d_r

  Look up the corresponding description for g_t in Table V, assign to d_t

end if

Compute the center of the reference bounding box $(x_r, y_r) \leftarrow \left(\frac{x_{r1}+x_{r2}}{2}, \frac{y_{r1}+y_{r2}}{2}\right)$

Compute the center of the target bounding box $(x_t, y_t) \leftarrow \left(\frac{x_{t1}+x_{t2}}{2}, \frac{y_{t1}+y_{t2}}{2}\right)$

$\Delta x \leftarrow x_t - x_r$

if $|\Delta x| > \gamma$ **then**

  if $\Delta x > 0$ **then**

    $d_{rel} \leftarrow$ "Right"

  else

    $d_{rel} \leftarrow$ "Left"

  end if

end if

return (d_r, d_t, d_{rel})

TABLE V: Mapping rules between grid sets and spatial descriptions

Grid set $grids$	Description d
$[1]$	Top-left
$[2]$	Top
$[3]$	Top-right
$[4]$	Left
$[5]$	Center
$[6]$	Right
$[7]$	Bottom-left
$[8]$	Bottom
$[9]$	Bottom-right
$[1, 2]$, $[1, 4]$	Top-left
$[2, 3]$, $[3, 6]$	Top-right
$[4, 7]$, $[7, 8]$	Bottom-left
$[6, 9]$, $[8, 9]$	Bottom-right
$[2, 5]$	Upper-center
$[4, 5]$	Left-center
$[5, 6]$	Right-center
$[5, 8]$	Lower-center
$[1, 2, 3]$	Top
$[4, 5, 6]$, $[2, 5, 8]$	Center
$[7, 8, 9]$	Bottom
$[1, 4, 7]$	Left
$[3, 6, 9]$	Right
$[1, 2, 4, 5]$	Upper-left-center
$[2, 3, 5, 6]$	Upper-right-center
$[4, 5, 7, 8]$	Lower-left-center
$[5, 6, 8, 9]$	Lower-right-center
$[1, 2, 3, 4, 5, 6]$	Upper-center
$[4, 5, 6, 7, 8, 9]$	Lower-center
$[1, 2, 4, 5, 7, 8]$	Left-center
$[2, 3, 5, 6, 8, 9]$	Right-center
$[1, 2, 3, 4, 5, 6, 7, 8, 9]$	Center

H. Depth Ordering Algorithm

Algorithm 2 Depth-Based Relative Position Description

Require: Depth values of Reference object $D_r = \{d_{r1}, d_{r2}, \ldots, d_{rn}\}$, Depth values of Target object $D_t = \{d_{t1}, d_{t2}, \ldots, d_{tm}\}$, Threshold γ

Ensure: Description of spatial depth ordering $desc$

Compute the average depth of the Reference object:

$$\bar{D}_r \leftarrow \frac{1}{n}\sum_{i=1}^{n} d_{ri}$$

Compute the average depth of the Target object:

$$\bar{D}_t \leftarrow \frac{1}{m}\sum_{i=1}^{m} d_{ti}$$

Compute depth difference: $\Delta D \leftarrow \bar{D}_r - \bar{D}_t$

if $\Delta D > \gamma$ **then**

  $desc$ ← "`<Target Object> is between the observer and the <Reference Object>.`"

  Optionally add:

  "`<Reference Object> is farther to the observer than the <Target Object>.`"

  "`<Target Object> is closer from the observer than the <Reference Object>.`"

else if $\Delta D < -\gamma$ **then**

  $desc$ ← "`<Reference Object> is between the observer and the <Target Object>.`"

  Optionally add:

  "`<Target Object> is farther to the observer than the <Reference Object>.`"

  "`<Reference Object> is closer from the observer than the <Target Object>.`"

end if

return $desc$

I. Structured Answer Format Design Techniques

Here, we will introduce the answer format design techniques in Fig. 3b

1) **Global Context (Overall Image Description):** Inspired by *Chain-of-Thought* [2] reasoning (e.g., having a language model write out detailed calculations improves accuracy on math tasks), we require the model to provide an overall description of the image at the beginning. This helps the model understand the global structure and context before proceeding to detailed reasoning.
2) **Absolute Directions `<Ref Abs Dir>` and `<Tgt Abs Dir>`:** Providing the absolute directions of the reference and target objects not only helps humans better understand their spatial relationships in the image, but also improves the accuracy of subsequent Egocentric and Allocentric view descriptions. These absolute directions are crucial for correctly inferring relative directions.
3) **Object Positions (`<Ref Pos>` and `<Tgt Pos>`):** Specifying object positions is particularly important when

the two objects are far apart in the image. For example, if a man is at the left boundary and a woman at the right boundary, it becomes difficult to describe their relationship naturally `<Ego Rel Dir>` without explicit position information. Including positions prevents such ambiguities.

4) **Egocentric Grounding:** Even though the Egocentric view is prompted initially with **"from the observer's point of view"**, we explicitly encourage the model to phrase answers like **"In the image, ..."** when resolving `<Ego Rel Dir>`. This ensures that reasoning is grounded in the observer's visual perspective.
5) **Depth Ordering:** If Step 3 (Depth Ordering Enhancement) is not implemented, ambiguities will arise. In the image, vertical relations (up–down) pose no ambiguity due to physical constraints. However, the remaining relations—front–back and left–right—are prone to ambiguity. Step 2 already resolves the 2D left–right issue, while Step 3 specifically addresses the 3D depth (front–back) relation to avoid such problems.Here, we adopt expressions such as **"The <Ref Obj> is between the observer and the <Tgt Obj>."** or **"The <Tgt Obj> is farther from the observer than the <Ref Obj>."** Instead of simply saying "in front" or "behind". For example, if the red bucket is in front of the blue bucket, the ambiguity arises because, even from an egocentric viewpoint, cultural backgrounds may differ: some people consider objects closer to the observer as being "in front", while others regard them as "behind".Another example is: "The man is facing the observer, the woman is facing away from the observer, and the man is in front of the woman." In this case, it is also unclear whether the man is actually closer to the observer or the woman is closer, leading to ambiguity.
6) **Allocentric Challenges:** There is currently no effective solution for allocentric view descriptions, and no model can accurately and reliably label them. Many objects cannot have their orientation directly defined from RGBD images—for example, a sofa, a monitor, or a microwave. Even if these objects inherently lack a clear orientation, humans tend to infer their direction based on usage context. Existing vision-language models (VLMs) perform poorly in this regard because they have not been trained with knowledge of "contextually defined orientation" during usage. This also highlights a key challenge addressed in this paper: not all cases can rely on tools to label data for improving VLM capabilities, which further validates the significance of our contribution.
7) **Simplifying Labeling:** During the labeling process, we tried to avoid overly complex scenarios—for example, when a person is "looking back." In such cases, it is difficult to define whether the person is facing toward or away from the observer, as interpretations can vary across individuals and cultures. Therefore, we excluded images of this type to keep the absolute direction as simple and consistent as possible. In the process of creating this dataset, similar issues may arise frequently. While some tasks may have clear definitions in spatial studies or cognitive psychology, here we rely solely on humans' intuitive understanding of direction. This remains a problem worth addressing in the future, with the potential for clearer definitions.
8) **Non-directional Objects:** Some objects (e.g., soccer balls, potted plants, stones, fruits) have no intrinsic orientation and are excluded from direction annotation.
9) **Geographic Directions:** Absolute geographic directions (e.g., north, south, east, west) are beyond the scope of this work.

J. Prompt for judging view awareness

Prompt: Determine whether the following response contains a Reference frame. Rule: If the response includes the phrase "from the perspective of ..." (or similar wording), answer "yes". Otherwise, answer "no". Only output "yes" or "no" —no extra text.

K. Additional Experiments (Examples of emergent reasoning)

(a) When we provide an input image together with a question about an object not present in the image, the Suf-DS model (highlighted in yellow) accurately identifies that the object is absent, whereas the Ins-DS model generates a response based solely on the image content and ignores the user's query.

(b) When we provide only textual input without any image, the Suf-DS model (highlighted in yellow) precisely answers the question correctly and explicitly notes the absence of image input. In contrast, the Ins-DS model produces hallucinated content that is entirely unrelated to both the image and the question.

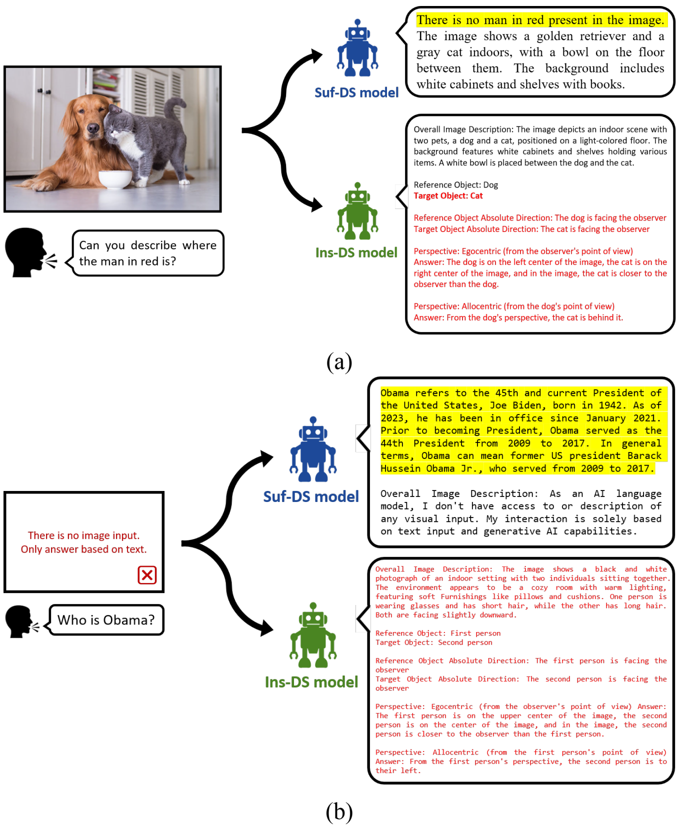


Fig. 10: Examples of emergent reasoning illustrate the image and semantic understanding capabilities of the Model trained on a sufficiently large dataset (Suf-DS model) and the Model trained on an insufficient dataset (Ins-DS model model).(a) Correctly responds that the image does not contain this object. (b) No image input.