AlloEgo-VLM: Resolving Allocentric and Egocentric Orientation Ambiguities in Visual-Language Model(s)

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. For instance, while a car is inherently non-directional, human usage scenarios typically imbue it with an assumed orientation. In natural language, spatial relationship descriptions frequently omit explicit reference frame specifications, leading to semantic ambiguity. Existing VLMs, due to insufficient annotation of reference frames and object orientations in training data, often produce inconsistent responses. Consider an image where a car is positioned on the left side facing left and a man stands on the right side facing the viewer: an egocentric perspective describes the man as "to the right of the car," whereas an allocentric perspective interprets him as "behind the car," highlighting semantic discrepancies arising from different reference frames. Such ambiguities can lead to erroneous decisions when robots rely on natural language for navigation and manipulation. To address this problem, 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 this representation, we constructed a dataset. By fine-tuning with QLoRA [1] , these spatial elements were integrated into a pre-trained 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 Models (VLMs), particularly in applications such as robot navigation, object manipulation, and human-robot interaction [2], [3]. However, spatial reasoning in natural language is inherently ambiguous due to the diversity of reference frames humans employ when describing relative positions. Cognitive psychology and spatial science have shown that humans often omit explicit reference frame specifications in conversation, relying instead on shared context or cultural conventions [4], [5], [6]. As shown 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 [7]. Such ambiguity is not only challenging for humans from different backgrounds but also poses a significant obstacle for VLMs, which often inherit this lack of explicit spatial grounding from their training data.

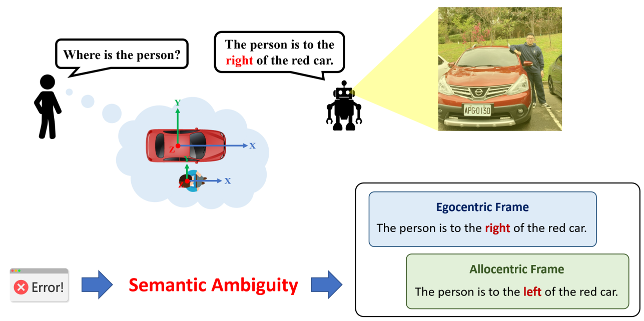


Fig. 1: Semantic Ambiguity Arising from Egocentric and Allocentric Frames.

Similar challenges have been observed in past works [8] on robot navigation [9], [10], Abstract Perspective Change [11], and video action prediction [12]. These studies have indeed introduced datasets or benchmarks with annotations under different reference frames, which provided valuable insights for spatial reasoning research. However, their formulations often require users to explicitly specify the reference frame, either directly or implicitly, in the query. For example, questions such as *"From the perspective of the viewer, where is the girl relative to the man?"* are commonly used to generate standard answers for training. In some cases, the evaluation is even presented as multiple-choice tasks, which do not fully reflect natural language usage in real-world settings. In everyday communication, humans rarely phrase their questions in such a verbose manner; instead, they simply ask *"Where is the girl relative to the man?"* without explicitly stating the perspective.

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 cat?"*, the system should automatically filter out inconsistent interpretations and produce the correct answer across different reference frames.

Existing VLMs are rarely trained with explicit annotations of reference frames and object orientations, resulting in inconsistent or even contradictory predictions when faced with spatial queries [13], [14]. This issue is particularly problematic in embodied AI robot systems, where incorrect spatial interpretation can lead to navigation errors, unsafe

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

manipulation, or task failures. While recent works have improved VLMs' ability to perform generic reasoning [15], [16], few have directly addressed the core problem of **spatial direction ambiguity** arising from missing reference frame information.

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) **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 [17] 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 3K carefully designed samples, outperforms existing state-of-the-art models in spatial orientation reasoning, demonstrating both efficiency and intelligence.

Furthermore, by leveraging Quantization and LoRA fine-tuning [18], [19], [1], the resulting models are lightweight enough for deployment on edge devices such as robots with limited VRAM. Our specialized inference strategy also enables the model to handle multi-turn dialogues effectively while maintaining strong performance on general vision-language tasks.

II. Related Work

A. Vision-Language Models and Spatial Reference Frames

The rapid development of large language models (LLMs) has demonstrated remarkable capabilities in text summarization, question answering, code generation, and multi-step reasoning [24], [25]. Instruction-tuning further aligns these models with human preferences [26], [27], and recent extensions have integrated multimodal capabilities, giving rise to vision-language models (VLMs) such as GPT-4o [28], LLaVA [29], and InstructBLIP [30]. These models show strong performance in interpreting and reasoning about visual content, motivating their application to spatial reasoning tasks where understanding object positions and relations is crucial.

However, prior work in vision-language reasoning has mainly emphasized geometric and relational properties of scenes, often overlooking the cognitive role of reference frames. Studies on hierarchical spatial structures [8], distance prediction [31], and object localization [32] achieve strong performance but fail to address how spatial descriptions shift across viewpoints. This issue is further reflected in benchmarks such as Spatial-Comfort [33] and ViewSpatial-Bench [10], which reveal that VLMs produce inconsistent outputs when reference frames are not explicitly encoded.

A key challenge lies in differentiating between egocentric and allocentric perspectives. Egocentric references, anchored to the observer' s viewpoint, are widely used in robotics for navigation and embodied perception [34], [35], whereas allocentric representations encode spatial relations independent of a particular observer, supporting tasks such as mapping and multi-agent coordination [36], [37]. Cognitive studies further show that humans flexibly switch between these perspectives depending on context and communicative efficiency [38], [39]. While recent VLM benchmarks attempt to incorporate perspective variation [40], they often require explicit perspective markers (e.g., "From the perspective of the viewer ......", "If I am standing by the stove and facing the dishwasher ......") or rely on multiple-choice formats, which deviate from natural language use.

Consequently, existing approaches do not fully resolve the challenge of reference-frame ambiguity in short, natural, and inherently ambiguous queries. This gap motivates our work, which focuses on enabling VLMs to robustly disambiguate spatial descriptions and produce consistent outputs across both egocentric and allocentric perspectives without requiring explicit perspective specification.

B. Training and Deployment Strategies for Spatial Grounding

Effective spatial grounding in vision-language models requires not only rich datasets but also carefully designed training and deployment strategies. The primary objective is not to eliminate the inherent ambiguity of natural language, but to ensure that models—or robotic agents powered by VLMs—avoid producing ambiguous or misleading outputs when interpreting spatial instructions [41]. In human communication, such ambiguities are often resolved through multi-turn dialogue, underscoring the need for models that can both interpret and clarify spatial references [42]. Supervised fine-tuning (SFT) plays a critical role in this process, as it aligns model behavior with human expectations and improves robustness in real-world interactions [26], [43]. However, general-purpose SFT alone is insufficient; specialized strategies that incorporate structured spatial supervision and explicit disambiguation mechanisms are necessary to effectively handle complex spatial scenes [44].

At the same time, deploying large-scale VLMs in real-world settings raises challenges of computational efficiency and resource constraints. Techniques such as LoRA (Low-Rank Adaptation) [18], quantization [45], [46], and knowledge distillation [47], [48] enable efficient fine-tuning and inference while preserving task performance, making them suitable for latency-sensitive applications such as autonomous navigation and human-robot interaction. More recently, modular and adapter-based architectures have been proposed to flexibly switch between base and fine-tuned

TABLE I: Comparison of Vision-Language Models and Datasets with Spatial Reasoning Capabilities. ✓= yes, $\times = no, \triangle = partial$

Method	Task Focus	Reference Frame Support	Multiple Choice Questions	Rigorous Question Input	Scale
CLEVR [20]	Visual reasoning (synthetic QA)	×	△	×	∼700k
GQA [21]	VQA + scene graph reasoning	×	△	×	∼22M
SpatialSense [22]	Pairwise spatial relation classification	×	×	×	∼11k
SPAR [23]	Spatial perception and reasoning	×	△	×	∼7M
Thinking in Space [9]	Video reasoning and memory	✓	✓	✓	∼5k
Perspective Aware [11]	Abstract perspective change	✓	×	✓	-
SPHERE [8]	Spatial perception and hierarchical evaluation of reasoning	✓	△	✓	∼2k
ViewSpatial-Bench [10]	Cross-viewpoint understanding and spatial reasoning	✓	×	✓	∼5k
VLMD4 [12]	Video spatial reasoning	✓	✓	×	∼1k
Ours	Reference frame reasoning	✓	×	×	∼4k

capabilities depending on task requirements and hardware availability [49], [50]. These approaches collectively facilitate scalable and efficient deployment of VLMs in edge scenarios without compromising their reasoning or multimodal understanding capabilities.

C. Datasets for Spatial Reasoning and Disambiguation

Several datasets have been proposed to advance spatial reasoning in vision-language models. CLEVR [20] provides synthetic images with compositional object arrangements and detailed spatial relationships, enabling models to perform visual question answering with controlled complexity. GQA [21] extends this by using real-world images annotated with structured scene graphs, supporting complex multi-step reasoning over visual scenes. SpatialSense [22] focuses specifically on natural images with annotated spatial relationships, emphasizing human-centric spatial semantics. More recently, the SPAR dataset [23] has been introduced to capture diverse spatial references and relational ambiguity, providing multiple valid interpretations for each spatial description.

Despite these advances, most existing datasets—including SPAR, CLEVR, GQA, and SpatialSense—do not explicitly model the impact of reference frames on spatial interpretation. As a result, vision-language models trained on them may still struggle with ambiguity when human descriptions omit reference-frame information, thereby limiting their reliability in tasks that demand precise spatial grounding [10]. A key challenge lies in differentiating between egocentric and allocentric perspectives: egocentric references, anchored to the observer’s viewpoint, are widely used in robotics for navigation and embodied perception [34], [35], whereas allocentric representations encode spatial relations independent of a particular observer, supporting tasks such as mapping and multi-agent coordination [36], [37]. Cognitive studies further show that humans flexibly switch between these perspectives depending on context and communicative efficiency [38], [39].

Recent benchmarks attempt to capture perspective variation [40], but they often require explicit perspective markers (e.g., “From the viewer’s perspective ...”) or adopt multiple-choice formats, which diverge from natural language use. Consequently, existing datasets do not fully resolve reference-frame ambiguity in short, natural, and inherently ambiguous queries.

Table I summarizes representative vision-language datasets and benchmarks for spatial reasoning, highlighting whether they support reference-frame reasoning, multiple-choice questions, or rigorous question inputs. This comparison illustrates the gap addressed by our work: we propose a structured spatial dataset explicitly designed to evaluate both egocentric and allocentric perspectives in realistic settings.

III. Structured Spatial Dataset generation

A. Image Acquisition and Preprocessing

As illustrated in Fig. 2, RGB images are first processed using an object detection model. In our pipeline, we employ YOLOv10 [51] Base for its fast inference speed. If the dataset already provides annotated object information, these annotations can be used directly. Initially, we filter images based on the number of objects, retaining only those containing approximately 2 to 6 distinct objects, which best align with the requirements of our task. Images that do not meet this criterion are discarded. Subsequently, each image is evaluated by a vision-language model (VLM), specifically GPT-4o [52], using an image screening prompt:

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”.

This secondary screening ensures that only images suitable for our dataset are retained. The selected images are drawn from multiple sources, including the GQA Dataset [21], SPAR Dataset [23], COCO Dataset [53], and NYU Depth Dataset V2 [54].

B. Assistant Response Generation (Answers)

After image preprocessing, each retained image is passed to a vision-language model (VLM), specifically GPT-4o [52],

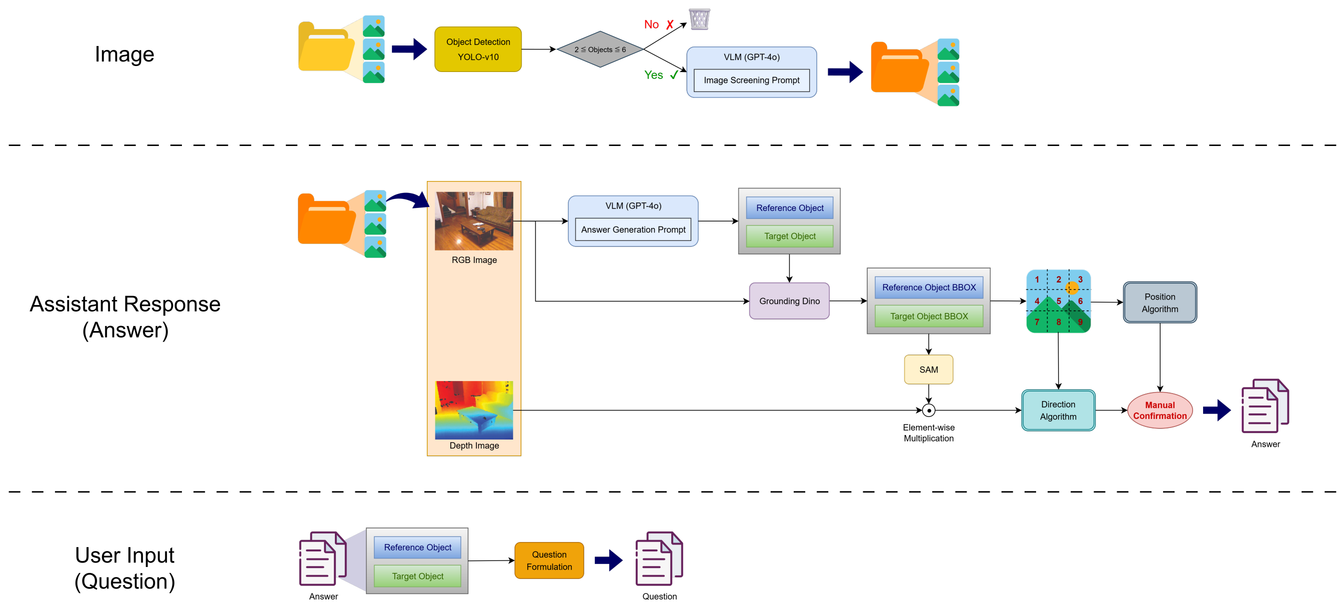


Fig. 2: Egocentric–Allocentric Dataset Pipeline. We construct triplets—Image, User Input, and Assistant Response—within a unified workflow to teach models how to disambiguate reference-frame ambiguity in spatial semantics.

using a carefully designed Textbook-level Answer Generation Prompt (see Fig. 3 and Fig. 4). This prompt enforces a detailed, standardized, and structured definition of the ground truth. For every response, the model produces: (1) a holistic scene description, (2) the reference object and its orientation, (3) the target object and its orientation, (4) the reference frame type, and (5) the relative positions defined within that frame. This design ensures that every answer is explicit, unambiguous, and suitable as a gold-standard annotation.

From each generated answer, the **<Reference Object>** and **<Target Object>** are extracted and paired with the original RGB image. These are then processed using Grounding DINO [55], which performs instance-level grounding. Unlike conventional object detection, which detects all objects at the category level, Grounding DINO [55]detects exactly one target instance specified by the description, producing **<Reference Object BBOX>** and **<Target Object BBOX>**.

Next, a Position Algorithm partitions the image into a 3×3 grid and determines the grid location of each bounding box. Based on its region, a natural language description is generated. For example, if the bounding box lies within regions 1, 2, 4, 5, 7, or 8, the algorithm outputs "the object is on the left-middle of the image," whereas if it lies exclusively in region 5, the output becomes "the object is in the center of the image." This ensures consistency in positional annotations.

Both **<Reference Object BBOX>** and **<Target Object BBOX>** are further processed by Segment Anything Model (SAM) [56], which extracts **<Reference Object Segmentation>** and **<Target Object Segmentation>**. The segmentation masks are then multiplied element-wise with the corresponding depth map, yielding **<Reference Object Depth>** and **<Target Object Depth>**.

Finally, a Direction Algorithm integrates depth information with the 3×3 positional encoding to produce directional relationships. For example: "In the image, the **<Target Object>** is to the left of the **<Reference Object>** and appears closer to the observer." This guarantees that both positional and depth-based relations are captured. The resulting structured responses are then manually confirmed to ensure correctness, forming the final ground truth annotations.

It is important to note that both Grounding DINO [55]and SAM [56]remain frozen throughout this process, serving purely as deterministic annotation tools rather than trainable components.

C. *User Input Collection (Questions)*

To complement the structured answers, we generate corresponding user queries based on the extracted <Reference Object BBOX> and <Target Object BBOX>. A set of seventeen question templates is predefined to capture diverse ways in which humans may inquire about spatial relationships (as illustrated in Fig. 5). These templates cover different phrasings, ranging from direct relational queries to more conversational descriptions, ensuring linguistic variability while maintaining semantic consistency. For clarity, two representative examples are shown below:

1) *"Where is the <Target Object> in relation to the <Reference Object>?"*
2) *"Can you describe where the <Target Object> is?"*

During dataset construction, one question is randomly sampled from this pool and paired with the corresponding assistant response. This strategy ensures that each data instance reflects natural variations in questioning style while remaining aligned with the ground truth spatial annotations.

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>.

Fig. 3: Standard Answer Format in the Dataset. The Assistant Response is constrained to a structured output, ensuring consistency across egocentric–allocentric tasks.

IV. Method

A. Multi-stage Iterative Training Framework

As illustrated in Fig. 6, we initially collected over 500 samples using the dataset generation method described in Section III. The process is then extended through an iterative pipeline:

- Step 1. Initial Fine-tuning. We apply QLoRA-based supervised fine-tuning (SFT) [1] on Qwen2.5-VL-7B [57], producing the Spatial Adapter that aligns the model with our structured spatial reasoning format.
- Step 2. Bootstrapped Data Generation. Unprocessed RGB images are fed into Qwen2.5-VL-7B [57]+ Spatial Adapter for inference. Due to the full-format SFT, the output structure is largely constrained, allowing the model to generate valid Assistant Responses (Answers) even without an explicit input question.
- Step 3. Question Generation. Each answer is paired with a corresponding User Input (Question) using the template pool defined in Fig. 5, thereby completing a consistent triplet: (Image, Answer, Question).
- Step 4. Human Verification. The generated triplets undergo secondary manual inspection to ensure both correctness and linguistic naturalness.
- Step 5. Iterative Refinement. The newly verified triplets, together with historical datasets, are used to further fine-tune Qwen2.5-VL-7B [57] + Spatial Adapter via QLoRA [1]. This cycle is repeated until convergence. In our setting, we perform five iterations (Step 1 $\rightarrow$ Step 5) to achieve stable spatial reasoning performance.

B. Question Spatial Classification

Due to the full-format SFT, the model tends to primarily respond to allocentric or egocentric spatial questions. To filter relevant tasks, we train a Spatial Classification model to determine whether a given User Input (Question) is related to spatial reasoning (see Fig. 7).

In our approach, the question text is first tokenized using a frozen DistilBERT tokenizer [58], producing the corresponding question tokens. These tokens are then passed through a frozen DistilBERT encoder [58] to obtain a question latent representation. Freezing both the tokenizer and encoder ensures that pre-trained language representations are preserved,

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

1. Based on the visual content of the image, please provide a comprehensive description of the entire scene. Your description
should include the following:
   - 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. Describe the absolute orientation of the <Reference Object>:
   - Do not compare it to any other object.
   - Only describe its own directional properties.
   - An object is facing the observer if its front side (e.g., face, headlights, screen) is visible and directly oriented
toward the viewer.
   - Use absolute terms like: "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. Describe the absolute orientation of the <Target Object>:
   - Do not compare it to any other object.
   - Only describe its own directional properties.
   - Use absolute terms like: "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".

4. Describe the 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.
   - All sections must be filled. If a description does not apply (e.g., no allocentric), explicitly state "No Allocentric."

   * Egocentric Description (Observer-Centered):
      I. Treat both the Reference Object and the Target Object as 2D bounding boxes in screen space.
      II. Think of the image as a nine-square grid. Describe their **individual screen positions** using the following
"standard screen positions" terms:
         - "upper left", "upper center", "upper right", "center left", "center", "center right", "lower left", "lower center",
"lower right".
      III. Then describe their **relative screen positions**, using terms like:
         - "<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>."
      IV. A depth relationship is considered visible if the image clearly shows one object occluding the other, or if
perspective and size cues suggest relative distance from the observer.
      V. If a depth relationship is visible, indicate which object is closer or farther from the observer using one of the
following sentence structures.  If not have depth relationship, don't reply:
         - "<Target Object> is between the observer and the <Reference Object>."
         - "<Reference Object> is closer to the observer than the <Target Object>."
         - "<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.

Use clear, objective, and neutral language. Avoid any subjective interpretations or emotional judgments. Respond strictly
using the specified format below:


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>.


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.
```

Fig. 4: Prompt for Assistant Response Generation. We employ an LLM (GPT-4o) with this prompt to generate preliminary Assistant Responses during dataset construction

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>?"

Fig. 5: User Input Collection. User inputs are generated from 17 predefined templates to provide consistent yet diverse spatial queries.

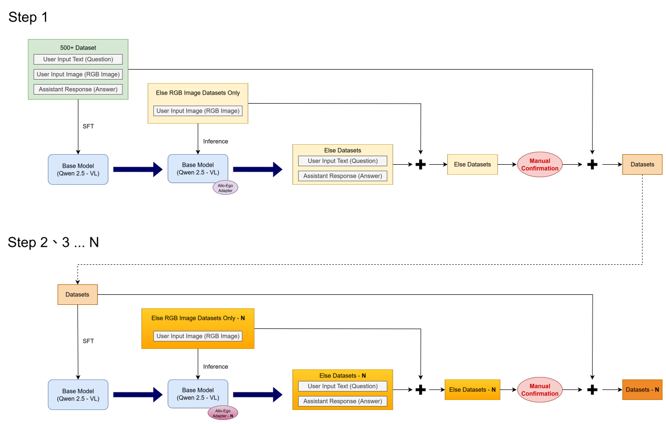


Fig. 6: Multi-Step Training Pipeline. Our training workflow, inspired by DPO, illustrates the stepwise procedure for model optimization.

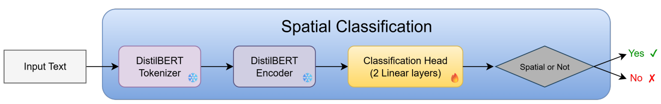


Fig. 7: Spatial Classification in the Model Pipeline. The module performs initial classification to determine if a sample falls within the scope of reference-frame disambiguation, serving as a pre-processing step for downstream reasoning.

reducing computational cost and improving generalization on small datasets. A classification head, consisting of two linear layers, is then trained on top of this latent representation to perform binary classification, determining whether the input question pertains to spatial-direction reasoning.

DistilBERT [58] was chosen for its efficiency and compact size while retaining strong semantic encoding capabilities. Its frozen embeddings provide robust and consistent text representations, making it well-suited for downstream classification tasks without requiring full fine-tuning.

The training dataset combines diverse question types collected from online sources, including mathematics, finance, and everyday queries, with the User Input (Question) generated as described in Section III. This mixture provides both positive and negative examples, enabling the classifier to robustly distinguish spatial tasks from other general-purpose questions.

C. Multi-round Dialogue Process

As illustrated in Fig. 8, during inference, each User Input (Question) is first processed by the Spatial Classification model (Fig. 7) to determine whether it corresponds to a spatial-direction reasoning task. If classified as spatial, the input is routed to Qwen2.5-VL-7B [57] + Spatial Adapter (fine-tuned as described in Section IV); otherwise, it is sent to the base Qwen2.5-VL-7B model [57].

The selected model then receives both the User Input (Question) and the associated RGB image, producing an Assistant Response that provides the requested information. The triplet—(User Input, RGB image, Assistant Response)

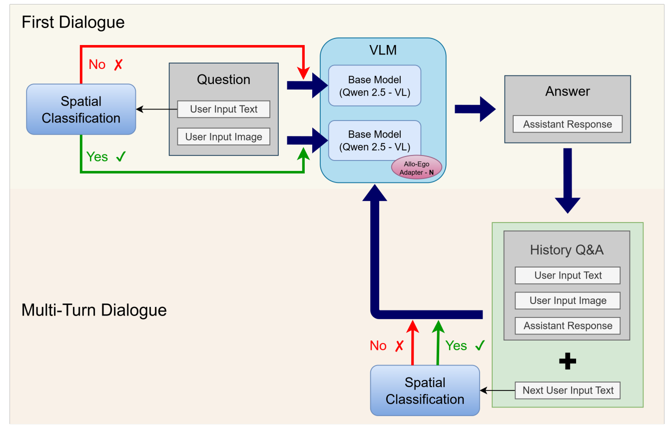


Fig. 8: Inference with Multi-Turn Dialogue. This workflow allows the model to iteratively interact and flexibly switch between egocentric–allocentric disambiguation and general reasoning tasks.

—is simultaneously recorded in the historical log to preserve context for subsequent interactions.

For each subsequent query, the next User Input (Question) is re-evaluated by the Spatial Classification model. The model then incorporates both the historical interaction records and the current input to generate a new Assistant Response, supporting coherent, context-aware, multi-turn spatial dialogue. This iterative procedure ensures that responses remain task-appropriate and consistent across consecutive conversational turns.

V. Experiment

A. Model Performance Evaluation

Table II 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** [52], **Llama 3.2-V** [59], **Gemma 3-V** [60], and variants of **Qwen 2.5-VL** [57]. The test dataset maintains the same structured format as described in Section III, ensuring consistency and fairness in evaluation.

Three evaluation settings are reported:

1) **Format-only prompt**: The model receives only the input question along with the expected output format, without additional guidance on spatial semantics or structured reasoning.
2) **Textbook-level prompt**: The model is provided with both the input question and the detailed prompt described in Fig. 4, which standardizes the response structure and emphasizes precise spatial reasoning, including object positions, reference frames, and directional relationships.
3) **Supervised Fine-Tuning (SFT)**: Models are trained using our iterative SFT pipeline described in Section IV, which includes structured dataset expansion, Spatial Adapter integration, and multi-turn spatial grounding. This approach ensures that the model con-

TABLE II: Model Performance Comparison. Performance of Qwen 2.5-VL 7B, Llama 3.2-V 11B, Gemma 3-V 4B, GPT-4o, and GPT-4o-min under Format-only prompts, Textbook-level prompts, and SFT scores. All User Inputs follow the template format in Fig. 5 and include both Reference Object and Target Object to ensure fair evaluation.

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

sistently produces unambiguous, contextually accurate, and task-aligned outputs.

The metrics reported are: **AD_RO** (Reference Object Absolute Direction), **AD_TO** (Target Object Absolute Direction), **Ego** (egocentric), and **Allo** (allocentric). Evaluation is performed automatically by **GPT-4o** [52], which compares the generated **Assistant Response** against the **Ground Truth Answer** using the scoring prompt illustrated in Fig. 9.

From Table II, several observations can be made:

- Models using **format-only prompts** achieve moderate performance, indicating that merely providing the question and output format is insufficient for complex spatial reasoning tasks.
- Incorporating the **textbook-level prompt** substantially improves performance across all metrics, demonstrating the importance of explicit guidance and structured output constraints.
- The full **SFT approach**, combining structured datasets, iterative fine-tuning, and Spatial Adapter integration, consistently achieves the highest scores in all categories, significantly outperforming GPT-4o and other SOTA models.

The superior performance of our approach can be attributed to several factors: (i) the dataset explicitly encodes reference objects, target objects, and directional relationships, reducing ambiguity; (ii) iterative SFT with human verification ensures that the model learns consistent spatial reasoning patterns; and (iii) the Spatial Adapter allows the base model to effectively generalize to diverse spatial configurations while maintaining structured output formats. Overall, these results highlight the effectiveness of our dataset and training methodology in enhancing VLMs' ability to perform precise spatial reasoning and generate contextually accurate, unambiguous outputs.

B. Question-Only Ambiguity Analysis

Table III reports the **Question-only Ambiguity Rate (AR)** for several state-of-the-art vision-language models. The ambiguity rate measures how frequently a model produces inconsistent or incorrect spatial responses when only the **User Input (Question)** is provided, without any explicit reference frame information.

```
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
   I. Assess how accurately the Reference Object Absolute Direction in the Assistant Response matches the Ground Truth Answer.
   II. Consider semantic correctness, direction consistency, and clarity.
2. Target Object Absolute Direction Score: X / 10
   I. Assess how accurately the Target Object Absolute Direction in the Assistant Response matches the Ground Truth Answer.
   II. Consider semantic correctness, direction consistency, and clarity.
3. Egocentric Answer Score: X / 10
   I. Assess how accurately the Egocentric answer in the Assistant Response matches the Ground Truth Answer.
   II. Consider whether the spatial relationship and positional details match.
4. Allocentric Answer Score: X / 10
   I. Assess how accurately the Allocentric answer in the Assistant Response matches the Ground Truth Answer.
   II. 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:
I. Whether meanings align
II. Whether any important details are missing or incorrect
III. Whether the response is misleading
IV. 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: ...
```

Fig. 9: Prompt for Automated Model Scoring. GPT-4o uses this prompt to assess the quality of model outputs, providing quantitative evaluation for Table II.

As shown, all evaluated models exhibit extremely high ambiguity rates: **Qwen 2.5-VL (7B)** [57] produces an AR of 94.57%, **Llama 3.2-V (11B)** [59]reaches 100%, and **GPT-4o** [52] 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 **Accuracy** column further reflects the models' ability to correctly answer spatial-direction reasoning questions under these constrained conditions, which remains low for most models, reinforcing the conclusion that question-only input is insufficient for reliable spatial reasoning.

C. Validation of Spatial Classification

As shown in Table IV, the Spatial Classification model achieves perfect scores across all validation metrics (Accuracy, Precision, Recall, and F1-score all 100%), demonstrating its ability to reliably distinguish spatial-direction reasoning tasks from other types of questions.

To provide further insight into the learned representations,

TABLE III: Question-Only Ambiguity Rate. Models are tested with question-only inputs, without restricting response format, to measure the occurrence of reference-frame ambiguity.

Method	Size	Question-only AR	Accuracy
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%

TABLE IV: Validation Results for Spatial Classification. We report accuracy, precision, recall, and F1-score for the Spatial Classification module described in Fig. 7, validating its effectiveness in filtering samples relevant to reference-frame disambiguation.

Validation Index	Spatial Classification
Accuracy	100%
Precision	100%
Recall	100%
F1-score	100%

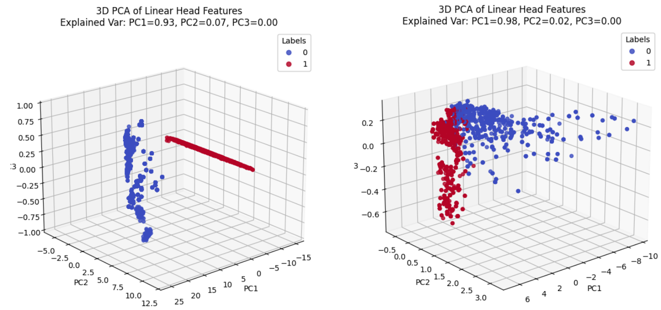
we visualize the last-layer linear M3 dataset features using 3D PCA [61]. Both subfigures in Fig. 10—(a) *Train set + Trained Model 3D PCA [61] Chart* and (b) *Test set + Trained Model 3D PCA [61]Chart*—show that the representations of spatial versus non-spatial inputs are well-separated. This confirms that the Spatial Classification model has learned highly discriminative latent features, effectively separating the classes in both training and unseen test data.

D. *Impact of Dataset Size on Model Behavior*

Table V summarizes the results of training Qwen2.5-VL-7B [57]and other vision-language models with different dataset sizes. When trained on the full dataset, Qwen2.5-VL-7B [57]demonstrates the ability to refuse to answer questions about irrelevant objects in an image while still producing accurate and contextually appropriate spatial responses. In contrast, when trained on only 10% of the dataset, the model tends to rigidly follow the prescribed response format, even for irrelevant objects, indicating reliance on format imitation due to insufficient training signals.

This behavior is not caused by overfitting. Evaluation on a held-out test set shows that models trained on the full dataset consistently outperform existing SOTA models such as GPT-4o [52]in spatial reasoning tasks. Models trained on 10% of the dataset perform roughly on par with GPT-4o [52], confirming that performance differences are due to dataset size and training signal rather than overfitting.

Interestingly, our dataset does not explicitly teach the model to refuse to answer, nor does it include additional general-purpose data. Under strong format supervision, one might expect the model to be "locked" into rigid format adherence. However, the results show that with sufficient data, the model emerges with the ability to judge when to apply prior knowledge and when to refuse irrelevant questions, indicating genuine learning of spatial semantics rather than mere format imitation.



(a) Train set + Trained Model 3D PCA Chart

(b) Test set + Trained Model 3D PCA Chart

Fig. 10: 3D PCA Projection of Model-Encoded Features. We apply PCA to the features extracted by the trained Spatial Classification module, showing that training and test samples are well separated, supporting the perfect validation metrics reported in Table IV.

TABLE V: Effect of Dataset Size on SFT Performance. SFT trained on the full dataset demonstrates broader generalization across tasks compared to SFT with 10% data, producing responses not restricted by the original format.✓ = success, × = failure.

Method	Refuse to answer	Text-only
Qwen 2.5-VL-7B (10% dataset)	×	×
Qwen 2.5-VL-7B (Full dataset)	✓	✓
Llama3.2-11B-Vision (Full dataset)	×	✓
Gemma3-4B-Vision (Full dataset)	×	×

Furthermore, the model demonstrates strong performance even on purely text-based questions, highlighting the robustness and generalizability of the learned representations. These findings underscore the value of our dataset and training methodology for enabling VLMs to acquire real knowledge and reasoning capabilities while exhibiting emergent behaviors under strong supervision.

For the other models, inference decoding parameters such as temperature, top-k, and top-p are set according to the recommended configurations provided by each model developer to ensure fair comparison. Llama3.2-11B-Vision [59]trained on the full dataset performs well on text-only questions but does not refuse to answer irrelevant objects, indicating weaker generalization and spatial reasoning ability compared to Qwen2.5-VL-7B [57]. On the other hand, Gemma3-4B-Vision [60], having fewer model parameters, performs poorly both in refusing irrelevant questions and in handling text-only inputs.

VI. Conclusions

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 [1] , 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 [52], Llama3.2-Vision [59], and Gemma3-Vision [60], in spatial reasoning tasks. Notably, Qwen2.5-VL-7B [57]trained on the full dataset exhibits emergent behavior, such as refusing to answer irrelevant questions, while maintaining robust performance on both image-based and text-only queries. This indicates that the model learns genuine spatial semantics rather than merely imitating response formats.

Our findings also highlight the importance of dataset scale and structured supervision: smaller training sets lead to over-reliance on format imitation, whereas larger, carefully annotated datasets facilitate emergent 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.

References

[1] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] J. Yang, S. Yang, A. W. Gupta, L. Fei-Fei, and S. Xie, "Thinking in space: How multimodal large language models see, remember, and recall spaces," *arXiv preprint arXiv:2412.14171*, 2024.

[10] 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.

[11] 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.

[12] 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.

[13] I. Stogiannidis, S. McDonagh, and S. A. Tsaftaris, "Mind the gap: Benchmarking spatial reasoning in vision-language models," *arXiv preprint arXiv:2503.19707*, 2025.

[14] 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.

[15] Y. Zha, K. Zhou, Y. Wu, Y. Wang, J. Feng, Z. Xu, S. Hao, Z. Liu, E. P. Xing, and Z. Hu, "Vision-g1: Towards general vision language reasoning with multi-domain data curation," *arXiv preprint arXiv:2508.12680*, 2025.

[16] T. Han, Y. Gao, *et al.*, "From diagnosis to improvement: Probing spatio-physical reasoning in vision language models," *arXiv preprint arXiv:2508.10770*, 2025.

[17] 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.

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[19] H. Xu, Z. Gan, Y. Cheng, and J. Liu, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[20] 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.

[21] 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.

[22] K. Yang, O. Russakovsky, and J. Deng, "Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition," in *International Conference on Computer Vision (ICCV)*, 2019.

[23] 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.

[24] 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.

[25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[26] 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.

[27] 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.

[28] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[29] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[30] 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.

[31] 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.

[32] A.-C. Cheng, H. Yin, Y. Fu, *et al.*, "Spatialrgpt: Grounded spatial reasoning in vision language models," *arXiv preprint arXiv:2406.01584*, 2024.
[33] 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.
[34] Y. Chen, Y. Xiong, T. Darrell, and X. Zhang, "Egocentric vision-based future vehicle localization for intelligent driving assistance systems," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.
[35] S. Gupta, J. C. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2616–2625.
[36] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MIT Press, 2005.
[37] E. C. Tolman, "Cognitive maps in rats and men," *Psychological Review*, vol. 55, no. 4, pp. 189–208, 1948.
[38] 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.
[39] A. Alexander, "Switching between egocentric and allocentric coordinates," Interview (Q&A) at Sainsbury Wellcome Centre Seminar, 2023.
[40] S. Cheng *et al.*, "Egothink: Evaluating first-person perspective thinking capability of vision–language models," in *Proceedings of CVPR 2024*, 2024.
[41] 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.
[42] J. Thomason, A. Padmakumar, J. Sinapov, N. Walker, Y. Jiang, H. Yedidsion, J. Hart, P. Stone, and R. Mooney, "Improving grounded natural language understanding through human-robot dialog," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6934–6941.
[43] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi, "Rethinking the role of demonstrations: What makes in-context learning work?" in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.
[44] H. Liu, Z. Zhang, and Y. J. Lee, "Improving instruction-following in vision-language models with conversational fine-tuning," *arXiv preprint arXiv:2304.03439*, 2023.
[45] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer, "Llm.int8(): 8-bit matrix multiplication for transformers at scale," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
[46] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," in *International Conference on Learning Representations (ICLR)*, 2023.
[47] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
[48] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision (IJCV)*, vol. 129, no. 6, pp. 1789–1819, 2021.
[49] J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, and I. Gurevych, "Adapterfusion: Non-destructive task composition for transfer learning," in *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.
[50] R. Karimi Mahabadi, S. Ruder, M. Dehghani, and J. Henderson, "Compacter: Efficient low-rank hypercomplex adapter layers," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
[51] 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.
[52] A. Hurst, A. Lerer, A. P. Goucher, and et al., "Gpt-4o system card," Tech. Rep., 2024.
[53] 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.
[54] 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.
[55] 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.
[56] 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.
[57] 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.
[58] 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.
[59] Meta AI, "Llama 3.2-v: Meta's multimodal vision-language model," *arXiv preprint arXiv:2409.12345*, 2024.
[60] Google DeepMind, "Gemma 3-v: Google's multimodal vision-language model," *arXiv preprint arXiv:2503.19786*, 2025.
[61] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.