Out of Sight, Not Out of Context? Egocentric Spatial Reasoning in VLMs Across Disjoint Frames

Sahithya Ravi[1*] **Gabriel Sarch**[2] **Vibhav Vineet**[3]
Andrew D. Wilson[3] **Balasaravanan Thoravi Kumaravel**[3]
[1] University of British Columbia [2] Carnegie Mellon University
[3] Microsoft Research, Redmond WA
sahiravi@cs.ubc.ca, gsarch@andrew.cmu.edu
{vivineet, awilson, bala.kumaravel}@microsoft.com

Abstract

An embodied AI assistant operating on egocentric video must integrate spatial cues across time – for instance, determining where an `object A`, glimpsed a few moments ago lies relative to an `object B` encountered later. We introduce DISJOINT-3DQA, a generative QA benchmark that evaluates this ability of VLMs by posing questions about object pairs that are not co-visible in the same frame. We evaluated seven state-of-the-art VLMs and found that models lag behind human performance by 28%, with steeper declines in accuracy (60% $\rightarrow$ 30 %) as the temporal gap widens. Our analysis further reveals that providing trajectories or bird's-eye-view projections to VLMs results in only marginal improvements, whereas providing oracle 3D coordinates leads to a substantial 20% performance increase. This highlights a core bottleneck of multi-frame VLMs in constructing and maintaining 3D scene representations over time from visual signals. DISJOINT-3DQA therefore sets a clear, measurable challenge for long-horizon spatial reasoning and aims to catalyze future research at the intersection of vision, language, and embodied AI. Code and data are available at https://github.com/sahithyaravi/DISJOINT-3DQA.

1 Introduction

We live in a three-dimensional world, and both humans and animals excel at building internal spatial representations that help them perceive, understand, and interact with their environments (Wang et al., 2002). For machines to act as capable embodied assistants, they too must be able to reason spatially: to infer where objects are, how they relate to one another, and how to navigate through space (Cheng et al., 2024; Chen et al., 2024; Cho et al., 2023). This is especially challenging in egocentric settings, where perception is anchored to a moving first-person viewpoint.

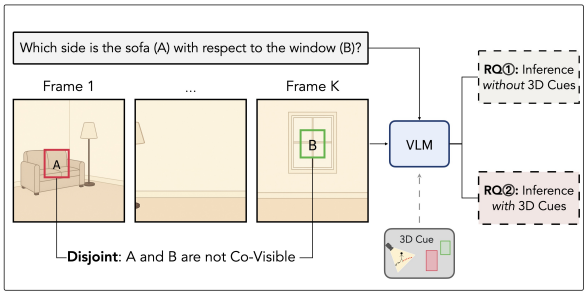


Figure 1: DISJOINT-3DQA: We focus on answering questions about spatial relationships when the objects are static but the camera is moving.

As the camera wearer moves, objects may enter and exit the field of view at different times, requiring models to reason across *temporally disjoint* observations. We term this setting **disjoint-frame spatial reasoning**: a model must accumulate geometric cues across time, 'mentally' reconstruct the scene, and then integrate across the cues to answer questions. While related to object permanence tracking (Tokmakov et al., 2021), our setting focuses on a harder subproblem, where objects do not co-occur, making spatial reasoning more challenging.

Figure 1 illustrates a typical example: the camera wearer first views a sofa (object A) in one part of the room and only much later encounters a window (object B) from a different viewpoint. The question *"Which side is the sofa (A) with respect to the window (B)?"* requires reasoning over temporally disjoint frames, where the two objects are not co-visible. This setup poses a fundamental challenge to current VLMs, which must infer spatial relationship among multiple frames. The core question we investigate is thus: *Can VLMs track and reason about spatial relationships when the relevant objects may not be co-visible?*

Recent work on embodied video understanding

[*]Work done during internship at MSR.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16135–16150
November 4-9, 2025

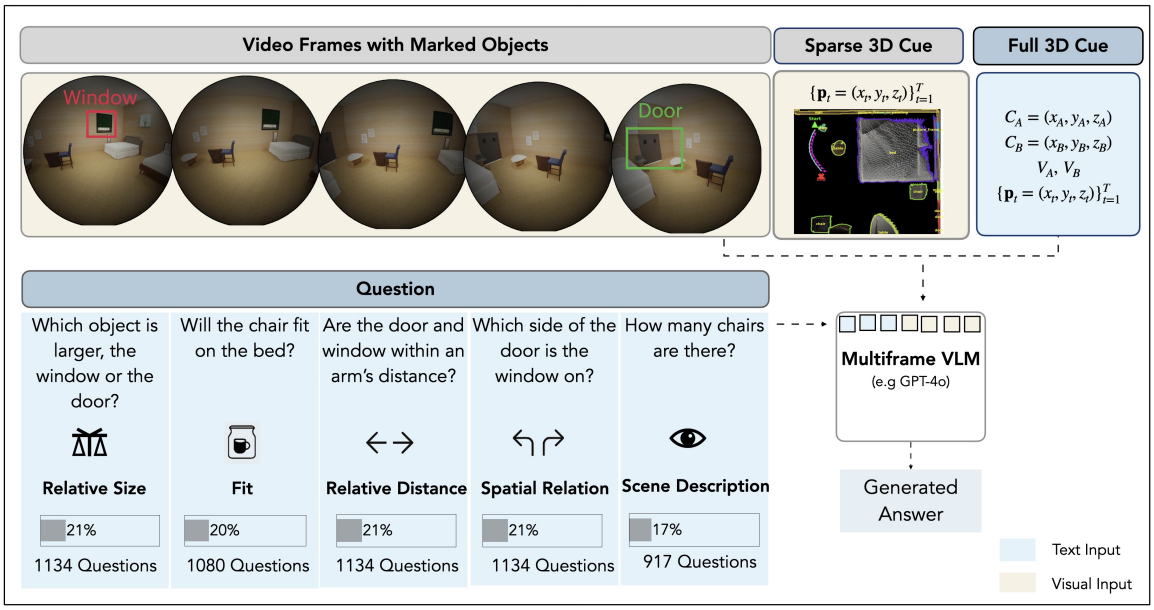


Figure 2: Demonstration of the evaluation setup of DISJOINT-3DQA. We provide the video frames with objects relevant to the question marked for visual grounding. Optionally, we evaluate models by providing explicit 3D cues using text or visual (Bird Eye View) inputs relevant to that question.

(Qiu et al., 2024; Amin and Rayz, 2024; Suglia et al., 2024; Majumdar et al., 2024; Cheng et al., 2024) and single-frame spatial reasoning (Kamath et al., 2023) show that language-based reasoning alone is insufficient for improving performance, even when objects are co-visible. The challenge becomes more acute in 3D settings: without explicit spatial priors, VLMs must infer scene geometry from raw pixels, often producing brittle or hallucinated maps of the scene (Yang et al., 2024b).

We introduce DISJOINT-3DQA[1], a generative QA benchmark that (i) poses spatial queries where relevant objects are not co-visible, compelling models to integrate information across frames and viewpoints and (ii) probe models with varying degree of explicit 3D scene information. Constructed from RGB-D egocentric recordings, DISJOINT-3DQA comprises 5,399 question–answer pairs spanning object-object relative direction, containment, and volumetric comparison, each requiring multi-frame spatial integration. We evaluate a spectrum of proprietary and open-source multi-frame VLMs including GPT-4o and, Qwen2.5VL and observe an overall lag behind human performance by 28%.

A natural hypothesis is that these models struggle because raw RGB frames provide limited geometric information, particularly in egocentric videos where relevant objects may never appear in the same frame due to continuous camera motion. We therefore augment the input with **explicit 3D context** in two forms: (i) *Text Trajectory Cue*, through textual camera trajectories, and (ii) *Top-down cue*, through bird's-eye view (BEV) renderings that depict the camera's path from object A to B. We refer to these as **sparse 3D cues**, because they expose only partial glimpses of the 3D scene geometry. While such cues lead to modest performance gains up to 2–3%, we observe significantly larger improvements exceeding 20%, when models are provided with **full 3D context**, including ground-truth object coordinates, volumes, and spatial metadata.

This contrast points to a central open challenge: enabling models to treat spatial cues not as isolated tokens, but as grounded elements of a coherent metric space. DISJOINT-3DQA thus motivates the need for models that can build and maintain internal 3D representations from sparse, temporally disjoint egocentric inputs, a capacity critical for robust spatial understanding in real-world environments.

2 DISJOINT-3DQA

Understanding grounded language begins with situating objects relative to one another, and to an embodied observer, within a coherent 3D world. Spatial reasoning for embodied agents typically arises in three settings: (i) when both camera and objects are static, (ii) when the camera moves through

[1]Dataset and code will be released upon publication.

a static environment, and (iii) when both camera and objects move. We focus on the second setting, common in egocentric video, where a moving agent observes a static scene from varying perspectives.

To evaluate spatial reasoning in such scenarios, we introduce DISJOINT-3DQA, a benchmark designed to test whether models can integrate information across temporally disjoint frames where relevant objects never appear together. This setup reflects real-world contexts like AR navigation or assistive robotics, where a complete spatial view is rarely available at once.

Figure 2 shows an example: a model must infer the spatial relation between a window seen early and a door seen much later. Such questions require integrating partial cues over time to recover spatial structure. DISJOINT-3DQA contains 5,399 question–answer pairs across 1,668 scenes and 856 unique object pairs, with an average of four diverse questions per scene. Motivated by foundational work in spatial cognition (Landau and Jackendoff, 1993), our questions span a range of reasoning types essential for embodied agents: relative direction, distance, size comparison, fit, and holistic scene understanding. Full dataset statistics are provided in Appendix A.1.

We build DISJOINT-3DQA using the Aria Synthetic Environments (ASE) dataset[2], a large-scale simulation of over 100,000 photorealistic indoor scenes out of which we randomly sample 1688 scenes. ASE offers rich 3D geometry, sensor metadata, and realistic object placements, making it ideal for spatial supervision. Its combination of realism, controllability, and scale enables high-precision question generation. Our construction pipeline is inherently **scalable** for two reasons: (a) it leverages the Aria Synthetic Environments, enabling controlled and repeatable data generation at scale, and (b) the steps in our pipeline are fully modular and parallelizable.

Goal. Given a pair of objects (A, B), the model must determine their spatial relation (e.g., left/right, size, relative distance) using only egocentric observations in which the objects appear separately. To construct such object pairs, we represent each video as a sequence of frames $\mathcal{F} = \{f_1, f_2, \ldots, f_T\}$, where each frame f_t contains a set of visible objects $\mathcal{O}_t$. For any object pair (A, B), we define their visibility spans as:

$$\mathcal{T}_A = \{t \mid A \in \mathcal{O}_t\}, \quad \mathcal{T}_B = \{t \mid B \in \mathcal{O}_t\}$$

We include (A, B) in the dataset only if $\mathcal{T}_A \cap \mathcal{T}_B = \emptyset$, ensuring that the two objects are never seen in the same frame i.e. not *co-visible*. This disjointness constraint forces models to accumulate spatial cues across non-overlapping views.

Ground-Truth Computation. To define ground-truth relations, we compute the position of object A relative to object B from the viewpoint of a frame where B is visible. Let $\mathbf{T}_B$ be the world-to-camera transform for that frame, and let $\mathbf{c}_A, \mathbf{c}_B \in \mathbb{R}^3$ denote the objects' centers in world coordinates. We transform them into the camera coordinate frame as follows: $\tilde{\mathbf{c}}_A = \mathbf{T}_B \cdot \mathbf{c}_A$ (object A in B's frame), $\tilde{\mathbf{c}}_B = \mathbf{T}_B \cdot \mathbf{c}_B = \mathbf{0}$ (object B at origin), and the relative offset is $\mathbf{d}_{AB} = \tilde{\mathbf{c}}_A - \tilde{\mathbf{c}}_B = \tilde{\mathbf{c}}_A$ where $T_B \in SE(3)$ is the world-to-camera extrinsic transform whose origin is fixed at the centre of object B and whose axes are aligned with the viewing direction of the image in which B is visible[3].

All spatial relations are derived from $\mathbf{d}_{AB}$, which encodes object A's position relative to B in the local frame. This anchors each question to a consistent egocentric perspective. This is consistent with the directional vector $\mathbf{d}_{AB}$ and the way spatial language is interpreted in egocentric video (e.g., "Is the ottoman to the left of the window?" is framed relative to where the *window* appears as shown in Figure 2). This design encourages the model to align its internal scene representation with the perspective of the current observation target and camera. To support diverse question types we also extract each object's volume $\mathbf{v}_A, \mathbf{v}_B$ and frame-wise instance maps detailing which objects appear in each frame.

Question Generation. For each object pair (A, B), and their respective reference frames (f_a, f_b), we have now derived all the 3D meta-data to answer spatial questions of different types. We then use predefined templates to come up with QA pairs using this meta data. We then provide this to GPT-4o to paraphrase it to more natural language question answer (QA) pairs. We provide the templates and prompts in Appendix A.2 and A.3.

Visual Grounding To ensure models attend to the correct object instances especially in scenes

[2] https://www.projectaria.com/datasets/ase/

[3] More details are discussed in Appendix A.9

containing multiple objects of the same type, we provide visual markers similar to Set of Marks (Yang et al., 2023). We project the 3D centers of objects A and B into their respective RGB frames using known camera parameters, and mark these projected centers with visual indicators (e.g., colored hollowed circles). This is demonstrated in Figure 2. We find this is an important factor influencing model performance, with *Marked* baselines outperforming the *Unmarked* counterparts significantly (§ 5.1).

2.1 Dataset Quality Evaluation

To verify data quality, we sampled 10% subset of the dataset and perform human evaluation on whether (i) the objects in the question are marked correctly in the video frames (yes/no) and (ii) the answer is accurate in the given context of video frames and the question (yes/no). We found that 99% of examples were correctly annotated with the appropriate object markers indicating the validity of our meta-data. Further, 96% of questions were relevant to the provided scene and 94% of answers were accurate [4].

3 Evaluating with 3D Cues

Can a vision-language model reason more effectively when provided with an explicit 3D scene representation, rather than relying on pixel-level inference alone? Visual object marking reduces referential ambiguity, but understanding spatial relationships between objects observed in disjoint frames often requires more than isolated 2D snapshots. To investigate this, we evaluate models under two types of 3D augmentations: **sparse** and **full 3D context**.

3.1 Sparse 3D Cues

Sparse 3D cues provide realistic, test-time signals that offer partial information about the scene's spatial layout. We introduce two forms:

Textual Trajectory Cue. We encode the camera trajectory as a sequence of positions $\mathbf{p}_t = (x_t, y_t z_t)$. This text-based representation reflects how the camera traverses the scene.

Top-down Cue. We generate a bird's-eye view (BEV) rendering of the scene. Built from RGB-D and instance segmentation data, the BEV provides a top-down visualization of the scene. It captures the spatial geometry that is not easily inferred from RGB frames alone. For each question, we render a targeted BEV image that highlights the relevant sub-trajectory, from the frame where object A appears to the frame where object B is visible. These visual cues are derived from the full 3D reconstruction but presented in a 2D RGB image that can be readily processed by modern VLMs. For example, in Figure 2, the BEV image shows the top down view of the scene, along with a trajectory involving the two objects in the question - ottoman and floor mat. Additional details on generation and prompting with BEV images are in Appendix A.4 and A.8.2.

3.2 Full 3D Context

In addition to evaluating models under sparse cues, we introduce an oracle setting where the model receives dense, ground-truth spatial metadata. This **full 3D context** includes the precise 3D coordinates of object centers in a global reference frame $C_A = (x_A, y_A, z_A)$ and $C_B = (x_B, y_B, z_B)$, as well as their physical dimensions or bounding box volumes V_A and V_B. This representation encodes the metric spatial relationships underlying the correct answer to each question. We use this setting to approximate an **upper bound on spatial reasoning performance**, isolating reasoning limitations from perceptual errors. While one might consider using predicted 3D detections instead, existing 3D detectors remain brittle, especially in egocentric video—due to occlusions, limited annotations, and poor generalization. Ground-truth cues thus serve as a clean scaffold to evaluate whether failures arise from missing information or from an inability to integrate and reason over spatial geometry.

4 Evaluation Setup

Metric. Our dataset follows a similar structure to OpenEQA (Majumdar et al., 2024), where each example consists of an open-ended question grounded in a visual or embodied context, with answers provided in natural language. Due to this alignment in task formulation and answer format (one line open-vocabulary answers), we adopt the LLM-Match metric proposed by OpenEQA to evaluate model predictions. This metric employs a large language model to rate the semantic similarity between predicted and reference answers on a 1–5 scale, providing a more reliable measure for open-ended, free-form QA than conventional string-matching methods. We normalize these scores to the [0, 1] range

[4] Appendix A.6 provides more details

and report them as percentages. Model and prompt for LLM-Match are detailed in Appendix A.7.

Models. We evaluate both closed-source and open-source VLMs. Closed-source models include GPT-4o accessed via public API. Open-source models include LLaVA-Next-Video (7B), LLava-Video (7B), InternVL (8B and 38B) and Qwen-VL (72B). For each model, we standardize the prompt format and provide a sequence of video frames along with the question. We prompt all models to provide a chain-of-thought (CoT) followed by the actual answer. Refer to Appendix A.8.1 for the prompts. These models are selected based on their strong performance on recent video reasoning benchmarks such as VideoMME(Fu et al., 2024) and their ability to process multi-frame inputs effectively.

Human Performance. We randomly sample a subset of 600 questions for estimating human performance on DISJOINT-3DQA. Three human evaluators independently answer each question, and their performance is evaluated using the same LLM-Match metric. Appendix A.6 provides further details on crowdsourcing.

5 Empirical Evaluation

We structure our evaluation around three core research questions designed to assess how well VLMs reason about spatial relationships in egocentric video, particularly when the objects in question never co-occur in the same frame.

① **RQ1: 2D-Only –** Can VLMs reason about spatial relations using only 2D egocentric video, and how does visually disambiguating object references affect performance?

② **RQ2: Effect of 3D Cues –** Does providing explicit 3D spatial cues, either linguistically or visually, enhance model performance?

③ **RQ3: Failure Modes –** What factors (e.g., object distance) make spatial reasoning particularly challenging for current models?

5.1 RQ1: 2D-Only — Does visual disambiguation improve spatial reasoning?

In the 2D-Only setting, we provide the VLMs with video frames and the question and investigate their performance out-of-the box *Unmarked* vs *Marked* . The *Marked* setting refers to the visual grounding setup described in 2, where objects relevant to the question are marked with red hollow circle to visually guide the models.

Figure 3a summarizes overall score and Figure 3b presents a breakdown of model performance across spatial categories. Closed-source models like **GPT-4o** exhibit notable spatial reasoning abilities out-of-the-box, reaching an accuracy of 62.88% even without visual marking of objects. With objects disambiguated, we see an improvement of nearly 3% over the *Unmarked* setting. Open-source models display greater sensitivity to visual prompting with markers, with improvements of approximately **7–9%** between the *Unmarked* and *Marked* settings. For instance, LLaVA-Video 7B, Qwen 2.5-72B, and InternVL3-8B all show substantial gains in response to object highlighting. These jumps suggest that even simple referential cues provide a strong inductive signal, helping models resolve ambiguous object references and reason more effectively in egocentric scenes.

As shown in Table 3a, humans achieve a normalized LLM-Match score of **93.96%**, outperforming all models by a wide margin. Notably, even the best-performing model—GPT-4o with visual markers—lags behind by over 28 percentage points. This performance gap persists across all spatial categories (Figure 3b). Beyond overall scores, the per-category analysis in Figure 2 reveals key trends. Most models perform well on categories such as **Relative Distance** and **Relative Size**, where spatial relationships are often visually salient and co-visible within frames and common notions understood in the language domain, such as a vase is likely smaller than a couch. However, all models, regardless of architecture or training data, struggle in more complex categories like **Size and Fit** and especially **Spatial Relationship**, which require multiframe integration.

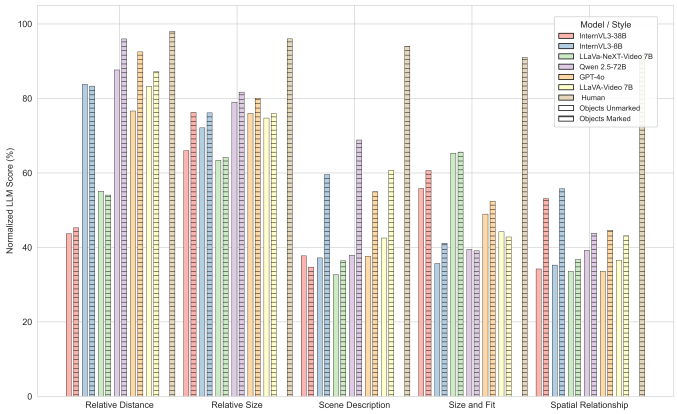
Takeaway: Explicit visual disambiguation significantly boosts model accuracy, particularly for open-source models. However, a substantial gap remains between the best-performing models and humans, highlighting limitations in current spatial reasoning and multi-frame integration.

5.2 RQ2: 3D Cues — Do 3D Cues Improve Reasoning in VLMs

Table 4a quantifies the effect of injecting 3D spatial information into VLMs. As described in § 3, we assess two settings: a *sparse augmentation* setting where test-time 3D cues (e.g., A→B bird's-eye

Model	Unmarked	Marked	Δ
Closed-Source Models			
GPT-4o	62.88	**65.60**	+2.72
Open-Source Models			
LLaVa-NeXT-Video 7B	48.44	49.92	+1.48
LLaVA-Video 7B	56.40	63.30	+6.90
InternVL3-8B	53.16	61.80	+8.64
InternVL3-38B	54.60	62.30	+7.70
Qwen 2.5-72B	55.06	**64.31**	**+9.25**
Human Performance	–	93.96	–

(a) Normalized LLM-Match (%) across models in two evaluation settings: **Unmarked** (no object highlighting) and **Marked** (referenced objects visually indicated).

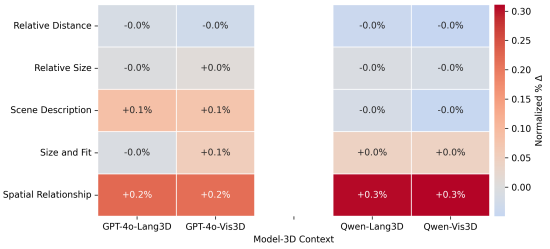


(b) Normalized LLM-Match (%) across spatial categories

Figure 3: RQ1: 2D-Only — Does visual disambiguation improve spatial reasoning: Comparison of LLM performance with and without visual object disambiguation. Left: overall accuracy across models. Right: performance by spatial category.

Modality	Input Type	Qwen 2.5-72B	GPT-4o
(1) Baselines			
–	No Marking	55.06	62.88
–	Visual Marking Only	64.31	65.60
(2) Sparse 3D Context (Test-time Cues)			
Text	3D Trajectory (Text)	66.21	67.60
Image	BEV (A→B Sub-Trajectory Only)	66.64	68.23
(3) Full 3D Context (Ground Truth Upperbound)			
Text	Direct Spatial Metadata (centers, volumes, trajectories)	-	83.2

(a) Effect of sparse and full 3D context on model performance. Normalized LLM Match (%) is reported.



(b) Δ from base to 3D-augmented prompts for GPT-4o and Qwen - Gains are seen in Scene Description and Spatial Relations.

Figure 4: RQ2: 3D Cues — Do 3D cues improve performance: Comparison of LLM performance with visual and linguistic 3D cues. Left: Overall performance with different cues. Right: Performance gained by spatial category.

view with sub-trajectories or language-based trajectories) are provided, and a *Full 3D context* setting with access to ground truth metadata (e.g., object volumes and positions). We compare it against a *baseline* with object markings only.

Across both GPT-4o and Qwen-72B, introducing sparse 3D cues consistently improves performance: BEV-based visual augmentation boosts GPT-4o by 2.6%, while linguistic descriptions yield slightly smaller gains. For Qwen, both visual and linguistic 3D inputs lead to ∼4% improvement.

Providing direct access to spatial metadata, such as object centers, volumes, and trajectories leads to a dramatic 18% jump in performance for GPT-4o. This upper bound reflects the advantage of structured geometry, where reasoning reduces to direct comparisons of object centers or volumes. Our results suggest that the bottleneck lies not in reasoning over 3D spatial inputs, but in constructing accurate 3D representations from sparse or 2D observations.

Category-wise Gains. To further dissect these improvements, Figure 4b (right) shows the normalized performance change (Δ) in five categories of spatial reasoning. The gains are most pronounced in *Scene Description* and *Spatial Relationship* questions, which demand a global understanding of object layout and egocentric traversal.

Takeaway: Realistic 3D cues offer modest gains especially for relational reasoning. The strong boost from ground-truth metadata highlights that the bottleneck lies in constructing accurate 3D representations from sparse cues or visual signals.

5.3 RQ3: Failure Modes

Distance Based Failure Modes. Figure 5 illustrates how model performance varies with the 3D Euclidean distance between object pairs. Across all models, accuracy declines as spatial separation increases. This trend is especially pronounced for

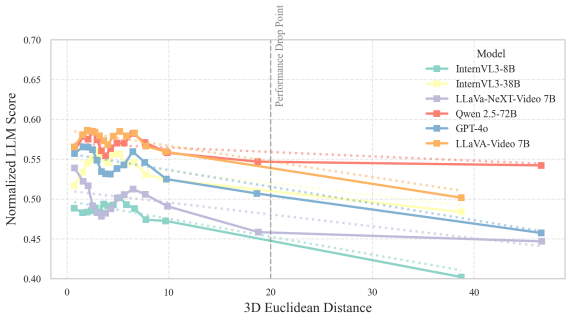


Figure 5: RQ3: Failure Modes: Model performance declines with increasing Euclidean distance between objects.

models like InternVL3-8B, which show sharper degradation. Larger models such as Qwen 2.5-72B and GPT-4o seem slightly more robust to moderate distances. However, performance drops significantly once the separation exceeds 20 metres, typically corresponding to objects located in different rooms.

Success vs. Failure of explicit cues Figure 6 compares the predictions of GPT-4o-based model under three visual grounding settings: (i) marked objects only, (ii) marked objects with egocentric trajectory, and (iii) marked objects with bird's-eye view (BEV) context. In the first row, the model must compare the size of the door and window across disjoint views. Without spatial grounding, GPT-4o misjudges the door as larger or expresses uncertainty. With BEV input, it correctly identifies the window as larger, matching ground truth, indicating that access to top-down scene views helps in understanding relative sizes. In the second example, the question concerns the directional spatial relation of whether the cabinet is to the right of the picture frame. Here, all three model variants answer wrongly, despite providing 3D context such as trajectories and the BEV.

Error analysis of Failure Modes of explicit cues. Across the full test split, adding BEV inputs yields a 2.5% improvement over the baseline for GPT-4o. We manually examined 50 examples where BEV maps led to improved performance. In 81% of these, the model's chain-of-thought (CoT) explicitly referenced the new cue e.g., "the BEV shows the window spans a wider area". However, an analysis of 50 failure cases with BEV inputs reveals several distinct error modes. In 18% of failures, the model mentioned the BEV but misinterpreted spatial relations, such as confusing directions or swapping the start and end of a trajectory. Another 17% ignored the BEV entirely without referencing the added spatial context. In 24% of cases, the model produced hallucinated geometry, making confident yet unfounded claims about object size, position, or visibility. We also observed that 15% of failures involved misalignment between egocentric views and BEV maps, where the model failed to correctly associate objects across views. These findings suggest that while BEV maps provide useful spatial context, their effectiveness depends on the model's ability to align, interpret, and selectively attend to the added modality.

6 Related Work

6.1 Egocentric 3D Understanding

Reasoning about space from an egocentric viewpoint is fundamental to embodied intelligence, enabling agents to navigate, manipulate, and interact with their environment (Ruggiero et al., 2009). Recent work has advanced 3D spatial understanding from egocentric inputs across several fronts. EgoGaussian (Zhang et al., 2024a) reconstructs static scenes using 3D Gaussian splatting from monocular egocentric videos. EgoSplat (Park et al., 2025) extends this approach with open-vocabulary capabilities and EgoSG (Zhang et al., 2024b) proposes building 3D scene graphs from egocentric footage. Large-scale egocentric datasets such as Ego4D (Grauman et al., 2022) and EPIC-KITCHENS (Damen et al., 2018, 2022) have catalyzed progress in this area, supporting tasks like action recognition, spatial localization, and object interaction.

6.2 Video Reasoning Benchmarks for VLMs

The growing capabilities of multimodal large language models (MLLMs) (Hurst et al., 2024; Team et al., 2024; Li et al., 2024a; Wang et al., 2024; Zhang et al., 2024c; Xue et al., 2024) have motivated the development of benchmarks to evaluate video understanding. Several recent efforts target third-person or general-purpose video reasoning, such as MVBench (Li et al., 2024b), VideoBench (Ning et al., 2023), TempCompass (Liu et al., 2024b), and Video-MME (Fu et al., 2024), which evaluate temporal ordering, event reasoning, and modality alignment. MotionBench (Hong et al., 2025) focuses on fine-grained motion understanding.

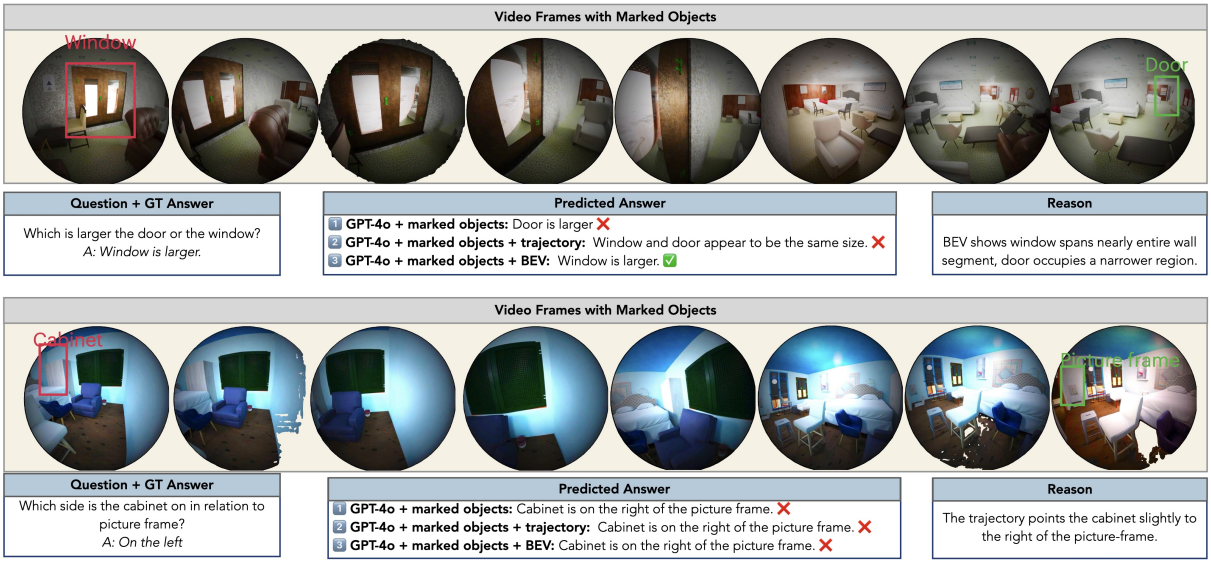


Figure 6: RQ3: Failure Modes. GPT-4o model predictions under different visual contexts: (i) marked objects, (ii) + trajectory, (iii) + BEV. Top: only BEV resolves spatial fit. Bottom: all models fail.

Dataset	Viewpoint	Reasoning Type	Object Co-visibility	Temporal Scope	3D Context	Task Format
SpatialRGPT-Bench (Cheng et al., 2024)	Mixed (Indoor/Outdoor/Simulated)	Spatial (2D/3D)	Varies	Single-frame	Yes	Gen
EmbSpatial-Bench (Newcombe, 2024)	Egocentric	Spatial (6 relations)	Mostly Yes	Multi-frame	No	MCQ
OpenEQA (Majumdar et al., 2024)	Egocentric	Spatial / Commonsense	Mostly Yes	Multi-frame	No	MCQ
VSI-Bench (Yang et al., 2024b)	Egocentric	Spatial (3D)	Yes	Multi-frame	No	MCQ
DISJOINT-3DQA (Ours)	Egocentric	Spatial (3D)	**No**	**Multi-frame**	**Optional**	Gen

Table 1: Comparison of recent video benchmarks closely related to DISJOINT-3DQA. Our benchmark, uniquely emphasizes spatial reasoning without object co-visibility and evaluates models by incorporating 3D scene structure.

In contrast, our work focuses on egocentric and embodied settings, where spatial reasoning is grounded in an agent-centric reference frame. Benchmarks such as EgoSchema (Mangalam et al., 2023) and EgoSpeak (Kim et al., 2025) evaluate language grounding and QA in egocentric contexts. OpenEQA (Majumdar et al., 2024) and VSI-Bench (Yang et al., 2024b) are most relevant to our focus on spatial reasoning. OpenEQA has spatial questions that often involve co-visible objects. VSI-Bench tasks involve generating cognitive maps from egocentric RGB inputs, but do not enforce object non-co-visibility and omit explicit 3D priors, effectively asking models to infer spatial structure without access to geometry. We explicitly enforce disjoint-frame spatial reasoning and optionally provide 3D structure, enabling a more controlled evaluation of spatial reasoning capabilities in VLMs. We show a comparison against spatial reasoning benchmarks in Table 1.

6.3 Spatio-Temporal Reasoning with MLLMs

Recent work has highlighted the limitations of MLLMs (Hurst et al., 2024; Team et al., 2024; Li et al., 2024a; Wang et al., 2024; Zhang et al., 2024c; Xue et al., 2024; Chinchure et al., 2024) in fine-grained spatial and temporal reasoning and worked on methods to improve them. Zhang et al. (2025) argue that architectural scaling alone is insufficient, advocating for spatially-aware objectives, structured supervision, and better positional encodings. More recent work has explored concrete strategies to improve VLM spatial reasoning. Liao et al. (2025) propose Group Relative Policy Optimization (GRPO), which fine-tunes VLMs using spatially grounded supervision and demonstrates substantial gains on spatial benchmarks. Parallel research explores methods such as coarse correspondence supervision (Liu et al., 2024a), and unified objectives for spatial understanding (Yang et al., 2024a; Chen et al., 2024; Cheng et al., 2024; Cai et al., 2024; Zhu et al., 2024).

7 Conclusion

We introduce DISJOINT-3DQA, a new benchmark designed to evaluate spatial reasoning in egocentric video where objects appear accors multiple frames. Through controlled experiments varying the availability of visual, textual, and 3D cues, we reveal key limitations of current vision-language models

(VLMs). Despite improvements from simple object marking and sparse 3D augmentations, models still struggle with tasks that require integrating information across disjoint frames. We hope our benchmark inspires the development of models that can internalize 3D priors, map 3D scenes, and robustly reason over complex egocentric environments.

Limitations

Sythetic Data. DISJOINT-3DQA is built using the Aria Synthetic Environments (ASE) dataset. While ASE is designed to be realistic and offers controllability and ground truth, the direct transferability of findings to the complexities and noise of real-world, unconstrained egocentric videos remains an open question.

Question Types. DISJOINT-3DQA spans a range of fundamental spatial relations. However, spatial reasoning includes other complex aspects or nuanced question types such as navigation (e.g, "How do I find my way back from the living room to the kitchen?"), layout (e.g, "How is the furniture organized in the room?") and compositional spatial reasoning (e.g.,"How do I clean my room?")

Level of "Sparsity" and Realism. The sparse cues used, while not providing the full 3D context – especially the BEV built from RGB-D and instance segmentation – are relatively processed and structured. In many real-world embodied AI scenarios, sparse 3D information might be noisier. It may not be representative of the more incomplete, or ambiguous nature of sparse 3D information often encountered in real-world scenarios.

3D Cues. The BEV map, while a useful abstraction, is a 2D projection of the 3D world. It simplifies the vertical dimension and can still have limitations in representing complex multi-level scenes or detailed object shapes from a top-down perspective.

Future work could address these limitations by incorporating real-world egocentric datasets, expanding the range of spatial question types, and exploring less structured or learned representations of 3D context that more closely mirror the noise and ambiguity found in practical scenarios.

References

Nadine Amin and Julia Rayz. 2024. Embodied language learning: Opportunities, challenges, and future directions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15369–15379, Bangkok, Thailand. Association for Computational Linguistics.

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. 2024. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*.

Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, and Leonid Sigal. 2024. Black swan: Abductive and defeasible video reasoning in unpredictable events. *Preprint*, arXiv:2412.05725.

Junmo Cho, Jaesik Yoon, and Sungjin Ahn. 2023. Spatially-aware transformers for embodied agents. In *ICLR*.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, Michael Wray, Hakan Bilen, Pauline Luc, Alireza Fathi, Martin Fiser, Daniel Goldberg, Kristen Grauman, Jiri Matas, and Cordelia Schmid. 2022. The EPIC-KITCHENS dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6055–6075.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. *Preprint*, arXiv:2501.02955.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Junhyeok Kim, Min Soo Kim, Jiwan Chung, Jungbin Cho, Jisoo Kim, Sungwoong Kim, Gyeongbo Sim, and Youngjae Yu. 2025. EgoSpeak: Learning when to speak for egocentric conversational agents in the wild. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2990–3005, Albuquerque, New Mexico. Association for Computational Linguistics.

Barbara Landau and Ray Jackendoff. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16:217–238.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*.

Zhenyi Liao, Kun Wang, Rongsheng Xu, Haoran Zhang, Guanlong Li, Peize Sun, and Jing Liu. 2025. Improved visual-spatial reasoning via r1-zero-like training. *arXiv preprint arXiv:2504.00883*.

Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. 2024a. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. *arXiv preprint arXiv:2408.00754*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024b. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772, Bangkok, Thailand. Association for Computational Linguistics.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, and 1 others. 2024. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*.

Nora S. Newcombe. 2024. *Spatial Cognition*. MIT Press. Https://oecs.mit.edu/pub/or750iar.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

Jin Park, Keren Li, and Kuk-Jin Yoon. 2025. Egosplat: Open-vocabulary egocentric scene understanding with language embedded 3d gaussian splatting. *arXiv preprint arXiv:2503.11345*.

Jielin Qiu, Mengdi Xu, William Han, Seungwhan Moon, and Ding Zhao. 2024. Embodied executable policy learning with language-based scene summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1896–1913, Mexico City, Mexico. Association for Computational Linguistics.

Gennaro Ruggiero, Tina Iachini, Francesco Ruotolo, and Vincenzo Paolo Senese. 2009. Spatial memory: The role of egocentric and allocentric frames of reference. In J. B. Thomas, editor, *Spatial Memory: Visuospatial Processes, Cognitive Performance and Developmental Effects*, 1st edition, pages 51–75. Nova Science Publishers, Hauppauge, NY.

Alessandro Suglia, Claudio Greco, Katie Baker, Jose L. Part, Ioannis Papaioannou, Arash Eshghi, Ioannis Konstas, and Oliver Lemon. 2024. AlanaVLM: A multimodal embodied AI foundation model for egocentric video understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11101–11122, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. 2021. Learning to track with object permanence. *Preprint*, arXiv:2103.14258.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Ranxiao Frances Wang, Elizabeth S. Spelke, Frances Ranxiao, and Wang. 2002. Human spatial representation: Insights from animals.

Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, and 1 others. 2024. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Jihan Yang, Runyu Ding, Ellis Brown, Xiaojuan Qi, and Saining Xie. 2024a. V-irl: Grounding virtual intelligence in real life. In *ECCV*.

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024b. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*.

Daiwei Zhang, Gengyan Li, Jiajie Li, Mickaël Bressieux, Otmar Hilliges, Marc Pollefeys, Luc Van Gool, and Xi Wang. 2024a. EgoGaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huanyu Zhang, Chengzu Li, Wenshan Wu, Shaoguang Mao, Yan Xia, Ivan Vulić, Zhang Zhang, Liang Wang, Tieniu Tan, and Furu Wei. 2025. A call for new recipes to enhance spatial reasoning in mllms. *arXiv preprint arXiv:2504.15037*.

Ronghan Zhang, Wencan Li, Siyuan Zhu, Li Zhang, and Yixin Zhu. 2024b. Egosg: Learning 3d scene graphs from egocentric rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2461–2471.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024c. Llava-next: A strong zero-shot video understanding model.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. 2024. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*.

A Appendix

A.1 Detailed Statistics of DISJOINT-3DQA

Table 2 summarizes key statistics of DISJOINT-3DQA. In figures 7, 8 we show the distributions of distances between object pairs in terms of 3D distance, number of frames. In Figure 9 we show the distribution of labels in object relationship questions.

Scenes	1668
Total QA pairs	5399
Unique object pairs (disjoint)	856
Avg. frames per question	12
Avg. questions per scene	4

Table 2: Summary statistics for DISJOINT-3DQA.

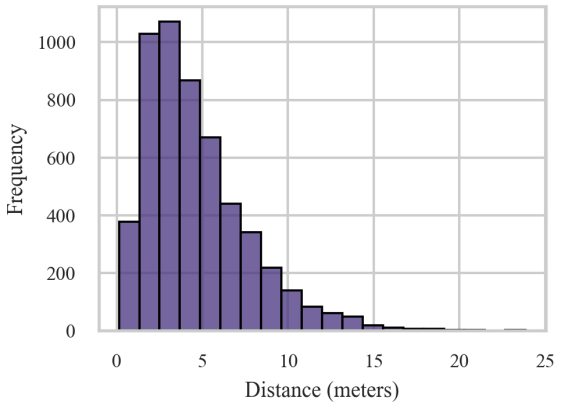


Figure 7: Distribution of spatial distances between object pairs in DISJOINT-3DQA. The majority of questions involve objects that are 2–6 meters apart, with a long tail extending up to 20 meters. This highlights the need for long-range spatial reasoning across frames.

A.2 Templates for DISJOINT-3DQA

To construct natural language questions, we design a set of templates aligned with core spatial reasoning capabilities. These templates reflect the underlying structure of each question type in the dataset and are instantiated with specific object labels (e.g., `object_a`, `object_b`) based on scene annotations. The resulting questions test models on relational reasoning, physical affordances, and basic object semantics in egocentric video. Some example templates are shown in 3.

A.3 Prompt for GPT-4o for Question Paraphrasing

To diversify question phrasing while preserving meaning, we use GPT-4o to generate paraphrased

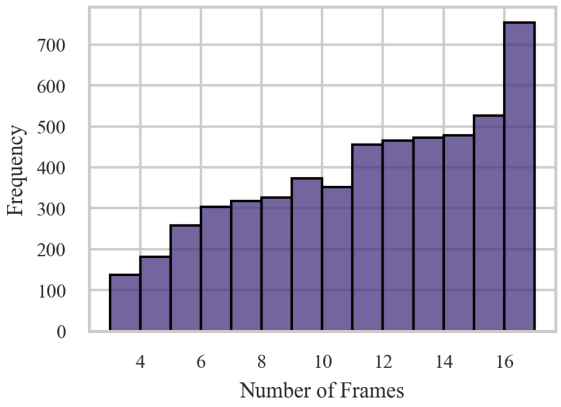


Figure 8: Distribution of the number of frames required to answer each question. Most questions span more than 10 frames, underscoring the need for multi-frame integration and memory over long temporal contexts.

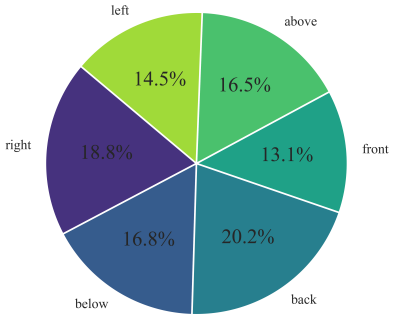


Figure 9: Distribution of directions for questions on spatial relationships.

variants of our base templates. The following system prompt is used:

```
Given a spatial question in
natural language, your task is
to rephrase it in a different and
natural manner while preserving
its meaning. The rephrased
question should not alter the
answer to the question. Do not
change the objects mentioned.
Avoid yes/no inversion.

Input: [Question] [Answer]
Output: [Paraphrased Question]
```

A.4 Top-Down BEV Rendering Algorithm

To visualize spatial configurations and egocentric camera motion in our dataset, we generate top-down bird's-eye view (BEV) maps using RGB-D

Category	Question Templates
Spatial Relationship	*Relative to the* `object_a`, *is the* `object_b` *on the left or right?*
	Is the `object_a` *vertically above or below the* `object_b`*?*
Relative Distance	*Are the* `object_a` *and the* `object_b` *within arm's reach of each other?*
	Can you touch both the `object_a` *and* `object_b` *from one spot?*
Size and Fit	*Which object is larger, the* `object_a` *or the* `object_b`*?*
	Can the `object_a` *fit on top of the* `object_b`*?*
	Can the `object_a` *be stacked on the* `object_b` *without falling?*
Scene Description	*List the unique objects in the scene?*
	What is the function of the `object_a` *in this scene?*

Table 3: Representative question templates across reasoning categories in DISJOINT-3DQA. Each template is grounded in egocentric visual context and instantiated with object pairs sampled from real scenes.

and instance segmentation data. The algorithm reconstructs a 3D point cloud from RGB-D frames using camera intrinsics and extrinsics, then projects this cloud to a global top-down map. Semantic instance regions are outlined, and camera poses are rendered as arrows and trajectories.

Listing 1: Core logic of BEV rendering using RGB-D data and camera poses.

```
for frame in frames:
    depth = load_depth(frame)
    rgb = load_rgb(frame)
    instance_map = load_instance_seg(
        frame)

    rays = compute_rays(intrinsics)
    points_cam = depth * rays
    points_world = transform_to_world(
        points_cam, extrinsics)

    instance_ids = instance_map[
        valid_pixels]
    color_overlay = assign_colors(
        instance_ids)

    # Accumulate points for rendering
    point_cloud.append(points_world)
    colors.append(color_overlay)

# Project to 2D grid
topdown_map = render_topdown(point_cloud
    , colors)

# Overlay camera trajectory
plot_trajectory(poses, topdown_map)
```

A.5 Example BEV Visualizations

Figure 10 shows three BEV examples from our dataset. Each map includes semantic instance regions (outlined by color), camera trajectory (cyan line), and start/end markers. These visualizations help disambiguate spatial relations across distant or occluded frames.

A.6 Crowdsourcing Protocol

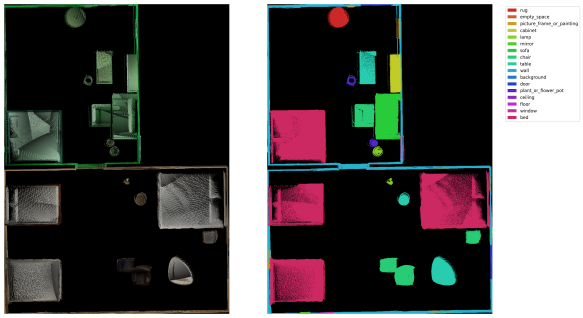
We use Amazon Mechanical Turk (AMT) to collect human performance baselines for DISJOINT-3DQA. For each evaluation type, every sample is independently annotated by three distinct workers to ensure reliability and diversity of responses. To ensure high-quality responses, we restrict access to workers with a HIT approval rate of at least 95% and more than 5,000 approved HITs. Workers are compensated at a rate of $9.99 for answering 15 questions, estimated based on pilot timing studies. Each annotation HIT includes clear instructions and example completions.

We report the average response accuracy across the three annotations per example. The template used for obtaining answers to the questions is provided in Figure 11 and template used for validatating DISJOINT-3DQAfollows a similar template, with YES/NO questions about the validity of the questions, answers and bounding boxes.

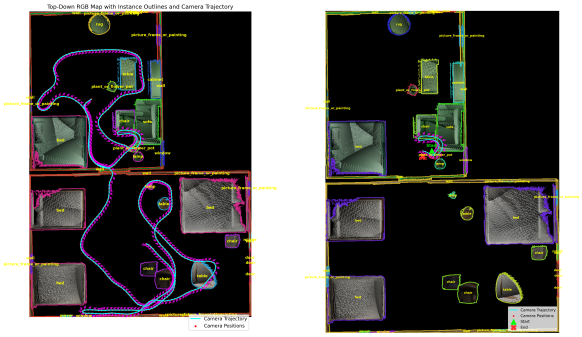
A.7 LLM-Match Prompt

We use the following prompt to obtain semantic similarity scores for model-generated answers, following the LLM-Match protocol introduced in OpenEQA (Majumdar et al., 2024). A large language model is prompted to rate the model's prediction on a scale from 1 to 5, based on its agreement with the reference and acceptable alternative answers:

You are an AI assistant who will help



(a) Top-down RGB map with object instance overlays.



Full trajectory Sub-trajectory (A → B)

(b) Egocentric camera path visualizations. Left: full trajectory. Right: question-specific sub-trajectory.

Figure 10: Top-down BEV maps for a scene in DISJOINT-3DQA. (a) Instance-wise spatial layout reconstructed from RGB-D frames. (b) Egocentric camera trajectories showing both global and question-specific paths.

me to evaluate the response given the question, the correct answer, and extra answers that are also correct. To mark a response, you should output a single integer between 1 and 5 (including 1, 5). 5 means that the response perfectly matches the answer or any of the extra answers. 1 means that the response is completely different from the answer and all of the extra answers.

Example 1: Question: Is it overcast? Answer: no Extra Answers: `["doesn't look like it", "no", "it's sunny"]` Response: yes Your mark: 1

Example 2: Question: Who is standing at the table? Answer: woman Extra Answers: `["a woman", "a lady", "woman"]` Response: Jessica Your mark: 3

Example 3: Question: Are there drapes to the right of the bed? Answer: yes Extra Answers: `["yes, there are drapes", "yeah", "the drapes are to the right of the king bed"]` Response: yes Your mark: 5

Your Turn: Question: question Answer: answer

A.8 QA Prompt

We provide GPT-4o with a series of egocentric frames and a natural language question. Below is the standardized prompt format used during inference.

Unmarked Objects.

You are a helpful assistant trained to answer spatial and visual questions based on egocentric video frames. Here are the Egocentric frames: Here is Frame 1 `[Image 1]`

Here is Frame 2 `[Image 2]`

...

Here is Frame N `[Image N]`

`[Question text]`

Please respond in the following format:

Reason: [Brief justification for your answer, 15–20 words]

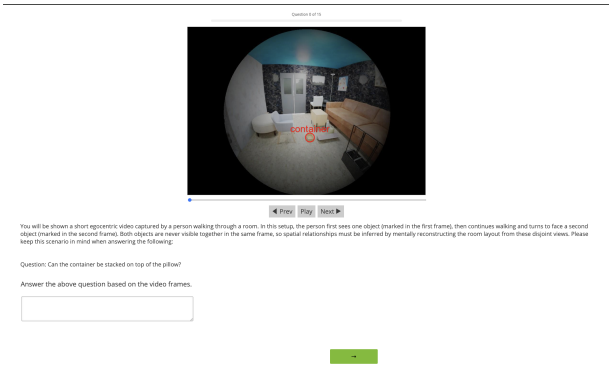


Figure 11: Template used for human evaluation

Answer: [Concise answer, max 15–20 words]

Unmarked Objects.

You are a helpful assistant trained to answer spatial and visual questions based on egocentric video frames.

Here are the Egocentric frames with highlighted objects in the first and last frame. The objects relevant to the question highlighted with a red hollow circle.:

Here is Frame 1 (object `A` and `B` marked)
`[Marked Image 1]`

...

Here is Frame N (object `A` and `B` marked)
`[Marked Image N]`

`[Question text]`

Please respond in the following format:

Reason: [Brief justification for your answer, 15–20 words]
Answer: [Concise answer, max 15–20 words]

A.8.1 Trajectory-Aware Prompt Format

For questions where camera pose information is available, we include 3D world coordinates per frame in the prompt. The system receives a list of egocentric frames along with their estimated (x, y, z) positions to help reason about spatial layout.

Example Prompt Structure:

You are a spatial reasoning assistant. Here is a sequence of egocentric frames. Use the visual evidence and the associated 3D camera positions to answer the spatial question.

Frame 1 with camera's 3D position (in meters): (2.53, -1.92, 1.50) [Frame image]

Frame 2 with camera's 3D position (in meters): (3.18, -1.21, 1.48) [Frame image]

Question: Is the bookshelf to the left of the couch?

Please respond in the following format: **Reason:** [15–20 word justification] **Answer:** [short answer, max 20 words]

A.8.2 Top-Down BEV Prompt Format

For certain questions, we supplement the egocentric video with a top-down reconstruction of the scene showing object instances and the camera trajectory.

Example Prompt Structure:

The image shows a top-down view of a 3D scene reconstructed from an egocentric video. The magenta arrows represent the camera's trajectory over time, based on frames relevant to the current question. The green triangle marks the starting camera position, and the blue X marks the ending position.

You need to focus on the part of the scene where the trajectory is marked to answer the question.

Question: Is the lamp behind the armchair?

Please respond in the following format: **Reason:** [15–20 word justification] **Answer:** [short answer, max 20 words]

A.9 World-to-Camera Transform T_B

Let $c_A, c_B \in \mathbb{R}^3$ denote the world-frame centers of objects A and B. We define a camera coordinate frame using the image where B is visible, placing its origin at c_B while preserving orientation. The resulting rigid-body transform is:

$$T_B = \begin{bmatrix} R_B & -R_B c_B \\ \mathbf{0}^\top & 1 \end{bmatrix}, \quad R_B \in \mathrm{SO}(3), \quad T_B \in SE(3). \tag{1}$$

This transform maps a world point $x_{\text{w}} \in \mathbb{R}^3$ to coordinates in the camera-B frame:

$$x_{\text{cam}} = T_B x_{\text{w}}, \quad x_{\text{cam}} \in \mathbb{R}^3. \tag{2}$$

By construction, T_B maps c_B to the origin:

$$T_B c_B = \mathbf{0}, \quad T_B \begin{bmatrix} c_B \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}. \tag{3}$$

Applying T_B to c_A gives the coordinates of object A in the camera-B frame:

$$\tilde{c}_A = T_B c_A = R_B c_A - R_B c_B. \tag{4}$$