
Towards Comprehensive Scene Understanding: Integrating First and Third-Person Views for LVLMs

Insu Lee^{1*}, Wooje Park^{1*}, Jaeyun Jang¹, Minyoung Noh¹,
Kyuhong Shim^{2†}, Byonghyo Shim^{1†}

¹Seoul National University, ²Sungkyunkwan University
{islee, wjpark, jyjang, mynoh, bshim}@islab.snu.ac.kr, khshim@skku.edu

Abstract

Large vision-language models (LVLMs) are increasingly deployed in interactive applications such as virtual and augmented reality, where a first-person (egocentric) view captured by head-mounted cameras serves as key input. While this view offers fine-grained cues about user attention and hand-object interactions, its narrow field of view and lack of global context often lead to failures on spatially or contextually demanding queries. To address this, we introduce a framework that augments egocentric inputs with third-person (exocentric) views, providing complementary information such as global scene layout and object visibility to LVLMs. We present E3VQA, the first benchmark for multi-view question answering with 4K high-quality question-answer pairs grounded in synchronized ego-exo image pairs. Additionally, we propose M3CoT, a training-free prompting technique that constructs a unified scene representation by integrating scene graphs from three complementary perspectives. M3CoT enables LVLMs to reason more effectively across views, yielding consistent performance gains (4.84% for GPT-4o and 5.94% for Gemini 2.0 Flash) over a recent CoT baseline. Our extensive evaluation reveals key strengths and limitations of LVLMs in multi-view reasoning and highlights the value of leveraging both egocentric and exocentric inputs. The dataset and source code are available at <https://github.com/Leeinsu1/Towards-Comprehensive-Scene-Understanding>.

1 Introduction

In recent years, large vision-language models (LVLMs) have received significant attention for their unprecedented performance in diverse tasks, including information retrieval, content generation, and multi-modal interaction [47, 18, 17, 13, 25, 16]. Their applications are now extended to interactive and immersive systems like virtual and augmented reality and embodied robotics [5, 24, 28]. Key characteristic of these applications is that images associated with the first-person view (a.k.a. egocentric view), captured by head-mounted cameras or smart glasses, are used as input. Although the first-person view is crucial in revealing the user’s intention, LVLMs might not interpret it well since most of the training datasets, often obtained from generic web images, are not acquired from a first-person perspective.

Recently, several approaches overcoming the scarcity of this so called egocentric data have been proposed. Notable ones include synthetic generation of egocentric data, incorporating large-scale egocentric datasets during the pre-training stage, and leveraging parameter-efficient fine-tuning on small-scale egocentric data [23, 33, 36]. While these approaches are effective to some extent, due to inherent constraints of the egocentric view, such as a limited field of view and a lack of global

* Co-first authors. † Co-corresponding authors.

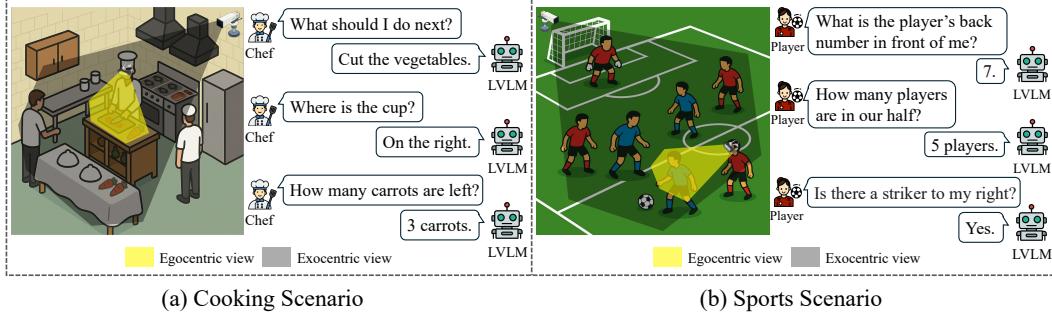


Figure 1: Conceptual illustration of example scenarios that require a joint understanding of egocentric (first-person) and exocentric (third-person) views. In each scenario, the first question can be answered using only the egocentric view, while the subsequent two questions require integrating information from both views. Yellow and gray overlays indicate egocentric and exocentric views, respectively.

context, LVLMs might not properly handle user queries requiring broader contextual understanding. Consider the scenario in Figure 1(a), where the user interacts with a visual assistant to make a food. When an egocentric image is provided as an input, LVLMs can readily answer the question “What should I do next?”, but may not provide the proper answer to questions such as “Where is the cup?” or “How many carrots are left?”, since the egocentric view has limited spatial coverage and thus fails to capture the comprehensive picture of the user environment. The moral of the story is that the egocentric view alone is not sufficient to handle the full range of user queries.

An aim of this paper is to propose a framework integrating multi-view information to enhance LVLMs’ capacity to understand the comprehensive context of a user’s surrounding environment. The core of our approach is to integrate images from the egocentric view and third-person view (a.k.a. exocentric view) to extract the merits of both. The exocentric view provides broader contextual cues, such as the user’s posture and the visibility of objects in blind spots, while the egocentric view offers information about user focus, including hand-object interactions and fine-grained object details. While integrating multiple viewpoints can help answer diverse user queries, it remains unclear whether conventional LVLMs can fully leverage them and really enhance the output quality. To answer the question systematically, we design 1) an ego-exo¹ multi-view question answering benchmark, referred to as the Ego-Exo Expanded Visual Question Answering (E3VQA) benchmark, and 2) a novel prompting technique called Multi-view, Multi-perspective, Multi-turn Chain-of-Thought prompting (M3CoT).

The purpose of E3VQA is to evaluate the ability of LVLMs to jointly perceive and reason when egocentric and exocentric views are provided. To this end, we meticulously curate a dataset comprising 4K question-answer pairs, each coupled with synchronized ego-exo images collected from the Ego-Exo4D dataset [12] covering four distinct tasks, including 1) action understanding, 2) object recognition, 3) spatial reasoning, and 4) numerical estimation. To fully leverage the complementary cues of ego-exo views, we further propose a novel prompting technique called M3CoT. Specifically, we construct a unified description of the scene (i.e., scene graphs) that guides the LVLM to understand the entire scene environment. This process consists of two main stages. First, we build three distinct scene graphs, each capturing the scene from a different perspective. The first graph is generated by simultaneously processing both ego and exo images, providing a holistic view of the scene while potentially overlooking fine-grained details. The second and third graphs are constructed by sequentially processing, either from ego to exo or from exo to ego, allowing the model to capture minute details. Second, we iteratively unify the three scene graphs, enabling each to progressively converge into a more robust and comprehensive representation. These unified scene graphs are used by LVLMs to generate the final responses.

We evaluate M3CoT on the E3VQA benchmark using state-of-the-art LVLMs, including GPT-4o [13] and Gemini 2.0 Flash [37], and observe a considerable gain in accuracy of 4.84% and 5.94% over the recent CoT method CCoT [32]. We also observe the noticeable gain (6.88% and 8.13%) on numerical reasoning questions, demonstrating our method’s effectiveness at integrating dual-view information.

In summary, our contributions are as follows:

¹Ego and exo denote the egocentric and exocentric, respectively.

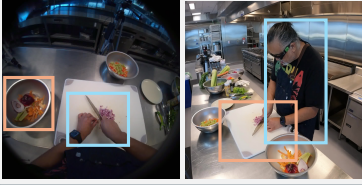
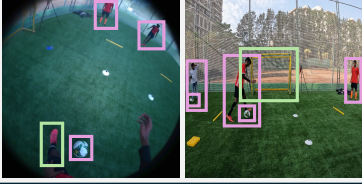
Pose & Action Perception <i>Questions about a person's physical state and movement, such as their poses, and actions</i> Q: What am I doing with my right hand? A) Chopping onions B) Grabbing a bowl C) Washing dishes D) Holding a fork Q: How is the person with glasses positioned? A) Sitting B) Kneeling C) Crouching D) Standing	Ego View Exo View 	Spatial Reasoning <i>Questions about the relative positioning and arrangement of objects or people</i> Q: Which object is to my left? A) Mixing bowl B) Plate C) Peeler D) Spice container Q: Which object is closest to the person? A) Mixing bowl B) Cutting board C) Spice container D) Plate
Object & Attribute Perception <i>Questions about identifying objects and their key features, like color or type</i> Q: What type of shoes am I wearing? A) Sneakers B) Cleats C) Sandals D) Flip-Flops Q: What color is the goalpost? A) Blue B) Green C) Yellow D) Red	Ego View Exo View 	Numerical Reasoning <i>Questions about counting elements, such as the number of people or objects in a scene</i> Q: How many people are there, including me? A) Two B) Three C) Four D) Five Q: How many soccer balls are on the field? A) One B) Two C) Three D) Four

Figure 2: Categories in the E3VQA benchmark. Each question is paired with ego-exo images and multiple-choice answers. The answers are highlighted in bold. The left part shows recognition categories, assessing the ability to focus on question-relevant parts. The right part shows reasoning categories, evaluating the ability to integrate information across views.

- We build the ego-exo multi-view VQA benchmark, E3VQA, consisting of 4K rigorously curated question-answer pairs with synchronized ego-exo image pairs. We construct E3VQA through a systemically designed pipeline to ensure that each instance evaluates the capabilities of LVLMs in integrating and reasoning across ego and exo views.
- To address the challenges posed by E3VQA, we propose M3CoT, a training-free prompting technique that combines scene graphs from three complementary perspectives into a unified scene graph. M3CoT improves the question answering accuracy on the E3VQA benchmark, achieving gains of 4.84% and 5.94% on two leading LVLMs, GPT-4o and Gemini 2.0 Flash.
- We perform a detailed analysis of leading LVLMs on E3VQA, uncovering their specific failure modes in multi-view reasoning and quantifying how egocentric and exocentric inputs affect performance.

2 E3VQA Benchmark

2.1 Motivation and Objectives

When compared to a single-image setting, LVLMs incorporating multi-view images face a number of challenges. First, the model needs to identify which image and what regions in an image are relevant to the question. Second, the model needs to filter out redundant content appearing in both views. Third, the model should deliberately extract ‘complementary’ cues from both views to generate a complete answer. Because a single image is provided per question for the conventional visual question answering (VQA) benchmark, they cannot evaluate multi-image reasoning capabilities.

To address this gap, we introduce E3VQA, a multiple-choice benchmark specifically designed for paired ego-exo images. Each question is accompanied by a set of plausible but incorrect options (i.e., *distractors*) that target typical failure patterns. These patterns include relying solely on one image (ego or exo), ignoring visual input altogether, or failing to merge complementary information from both views. These carefully crafted distractors enable E3VQA to precisely evaluate a model’s ability to reason across ego-exo image pairs.

2.2 E3VQA Composition

We organize E3VQA into four categories: pose and action perception, object and attribute perception, numerical reasoning, and spatial reasoning, to encompass a wide range of real-world scenarios (see Figure 2). Each category contains 1,000 question-answer (QA) pairs, evenly divided between egocentric and exocentric questions (e.g., “What am I doing?” vs. “What is the person doing?”), which supports the evaluation of the model’s generalization capability to diverse forms of user queries.

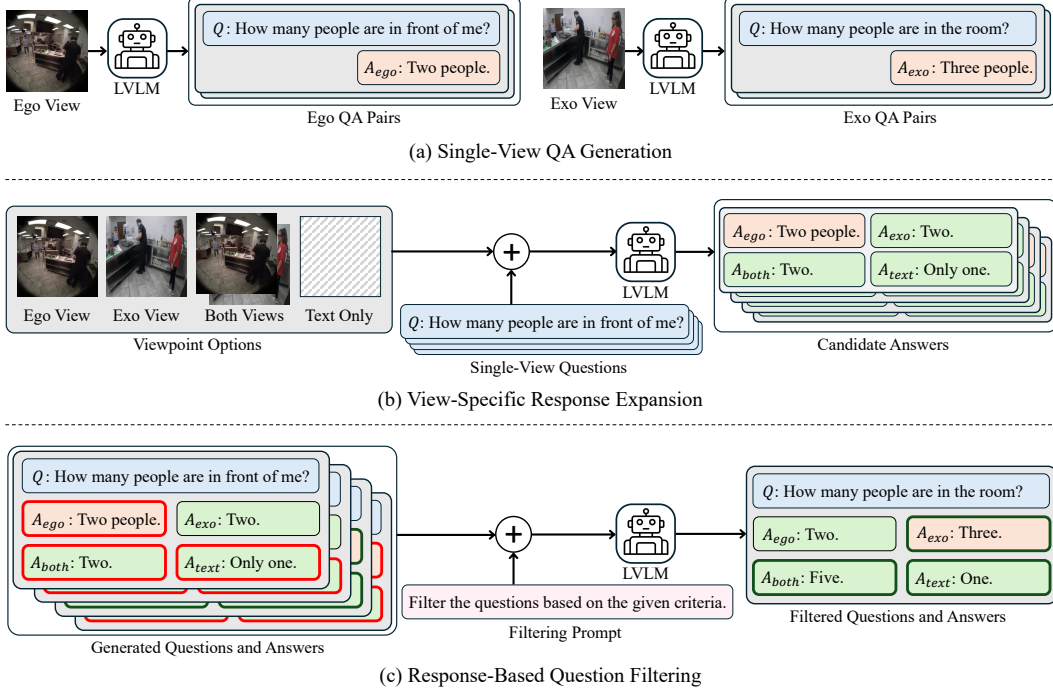


Figure 3: Overview of the E3VQA benchmark’s three-step automated QA generation pipeline: (a) single-view QA generation step, (b) view-specific response expansion step, and (c) response-based question filtering step.

To solve the questions, the model must identify the relevant object or person in both views and align their features across the two images. Variations in viewpoints and fields of view can make the same element look distorted, partially occluded, or differently scaled, making it difficult to establish correspondences and integrate visual cues. Detailed explanations of the challenges for each category and dataset statistics are provided in Appendices A.1 and A.2.

2.3 Dataset Construction Pipeline

2.3.1 Source Data Pre-Processing

The E3VQA benchmark is constructed utilizing the large-scale synchronized ego-exo dataset, Ego-Exo4D [12], which is composed of diverse user interactions (e.g., cooking, bike repair, soccer) filmed in various environments and countries. To ensure diversity within the dataset, video clips are uniformly sampled from the Ego-Exo4D dataset with respect to both user activities and recording locations. Each selected video clip is downsampled to 1 frame per second, and 8 frames are uniformly sampled from the downsampled clip. As a result, 4,600 ego-exo image pairs are obtained from the 575 video clips. Note that all video clips are selected from the test split to prevent any potential dataset contamination.

2.3.2 Automated QA Generation

We introduce a three-step automated QA generation pipeline designed to minimize human effort and improve scalability (illustrated in Figure 3). Throughout the entire process, we utilize GPT-4o [13], a powerful off-the-shelf LVLm.

Step 1: Single-View QA Generation We begin by generating QA pairs independently from either the ego or exo images, under the assumption that recent LVLms show a stronger understanding of single images than of multi-view inputs. Specifically, we instruct the model to generate Ego QA pairs (e.g., Q : What am I doing?, A_{ego} : Pouring water) from the ego images, and Exo QA pairs (e.g., Q :

What is the person doing?, A_{exo} : Stirring eggs) from the exo images. This design ensures that the generated questions align with the visual characteristics of each view: the ego image enables clear identification of the ‘I’, while the exo image provides a broader field of view and richer contextual information about ‘the person’. For each image, we generate three QA pairs per category to balance diversity and relevance; generating more than three often results in questions that are not visually grounded. As a result, this process yields a total of 110,400 single-view QA pairs.

Step 2: View-Specific Response Expansion In this step, we generate diverse candidate answers by presenting the model with the same question under four distinct input conditions: 1) ego view only, 2) exo view only, 3) both ego and exo views, and 4) text only (no visual input). The resulting answers are denoted by A_{ego} , A_{exo} , A_{both} , and A_{text} , respectively. These responses serve two key purposes in subsequent steps: 1) they are used as criteria for identifying low-quality QAs during the filtering stage, and 2) they function as hard candidate options for constructing multiple-choice questions in the human verification stage.

Step 3: Response-Based Question Filtering A large proportion of the questions generated in the previous step are either too easy or disqualified for multi-view question answering. To filter out such questions at scale, we introduce a response-based question filtering strategy that uses the initial answer from Step 1 (denoted as A_{init} ; either A_{ego} or A_{exo}) as a reference. First, we discard questions where A_{text} matches A_{init} , since this indicates that the question can be answered without any visual input. Second, we remove questions for which A_{both} matches A_{init} , which indicates that the answer remains unchanged even when both ego and exo views are provided, and therefore implies that multi-view reasoning is unnecessary. Applying these two criteria retains only those questions that cannot be answered without integrating both views. As a result, approximately 78.5% of the initial questions are filtered out, leaving a set of 23,694 challenging, multi-view QA samples.

2.3.3 Human Verification

Following automated QA generation, we perform a human verification with four expert annotators. The experts thoroughly review all questions, discarding unclear or low-quality questions. They also carefully craft the answer options, leveraging the responses generated in the previous step: A_{ego} , A_{exo} , A_{both} , and A_{text} . This process results in the final E3VQA dataset, comprising 4,000 high-quality QA pairs, representing just 3.6% of the original 110,400 samples after filtering and refinement. Please refer to Appendix A.3 for additional details.

3 M3CoT: Multi-Perspective Scene Understanding

3.1 Multi-Perspective Scene Graph Generation

In our proposed ego-exo multi-image question answering scenario, we expect the LVLM to generate the most appropriate answer given a query Q and a pair of ego and exo images $\mathbf{I} = \{I_{ego}, I_{exo}\}$. To help the model understand ego and exo images comprehensively, we employ a multi-perspective scene graph generation approach. Specifically, three instances of an LVLM act as distinct agents, denoted as $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$, each generating a scene graph from a different perspective as follows:

- **Ego&Exo:** Agent $\mathcal{F}_1$ generates a joint scene graph S_1 in a single step by simultaneously processing both I_{ego} and I_{exo} as input.
- **Ego2Exo:** Agent $\mathcal{F}_2$ first generates a scene graph using only I_{ego} , which is then sequentially expanded by incorporating information from I_{exo} to generate scene graph S_2 .
- **Exo2Ego:** Agent $\mathcal{F}_3$ follows the reverse approach, generating a scene graph based solely on I_{exo} and subsequently supplementing it with I_{ego} to generate scene graph S_3 .

Together, the three agents can capture both view-specific details and holistic scene context from complementary perspectives. For each perspective, prompts are carefully designed to capture the complementary information present in the ego-exo images. Please refer to Appendix E.2 for the complete prompts.

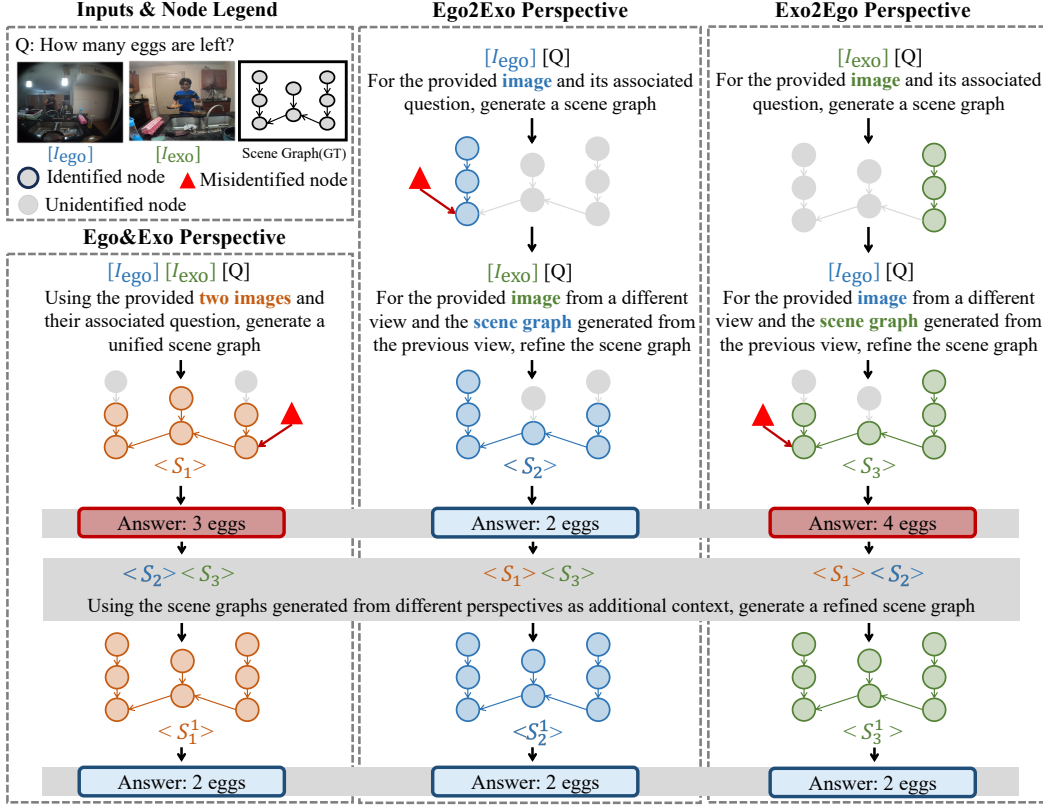


Figure 4: Overview of the M3CoT method. **Left:** Scene graph generation process from the Ego&Exo perspectives. **Center:** Scene graph generation process from the Ego2Exo perspective. **Right:** Scene graph generation process from the Exo2Ego perspective. Scene graphs from each perspective are merged to complement missing objects and relations, enabling the model to perform coherent reasoning and answer generation.

3.2 Iterative Multi-Agent Scene Graph Refinement

To further refine each scene graph S_i (for $i = 1, 2, 3$) generated by agent $\mathcal{F}_i$, we iteratively incorporate information from the other two agents. At iteration t , agent $\mathcal{F}_i$ takes the other two scene graphs (e.g., S_2^{t-1} and S_3^{t-1} for $\mathcal{F}_1$) to examine their objects, attributes, and relationships, and then adjusts S_i^{t-1} to S_i^t to better align with both I_{ego} and I_{exo} . Here, S_i^t denotes the scene graph S_i after the t -th update. By leveraging complementary information from multiple perspectives, this process improves both the accuracy and completeness of each agent’s scene graph. At each iteration step t , every agent $\mathcal{F}_i$ generates an answer to the question, conditioned on S_i^t . We then aggregate the agents’ responses via majority voting. If a consensus is achieved, we accept the majority answer and terminate the process. If the agents’ answers remain inconsistent after a fixed number of iterations, the final answer is selected from the response of $\mathcal{F}_1$. This iterative loop yields progressively richer scene representations and promotes convergence among the agents’ answers.

4 Experimental Results

4.1 LVLMS’ Performance on E3VQA

To assess ego-exo multi-image reasoning capabilities, we evaluate five closed-source and nine open-source LVLMS on the E3VQA benchmark using their default configurations. Detailed model specifications, experimental settings, as well as system and user prompt templates are provided in Appendices B.1 and E.2.

Table 1: Performance comparison of recent closed and open-source models on the E3VQA benchmark.

LVLMs	Pose & Action		Object & Attribute		Numerical		Spatial		Avg.
	Ego	Exo	Ego	Exo	Ego	Exo	Ego	Exo	
Closed-Source									
GPT-4o [13]	49.87 $\pm$ 1.10	63.47 $\pm$ 0.81	72.47 $\pm$ 0.12	77.00 $\pm$ 0.40	48.47 $\pm$ 1.14	57.67 $\pm$ 0.12	61.13 $\pm$ 0.99	57.13 $\pm$ 0.12	60.90
GPT-4o mini [13]	41.33 $\pm$ 0.31	49.00 $\pm$ 0.40	66.07 $\pm$ 0.12	71.00 $\pm$ 0.00	35.80 $\pm$ 0.35	44.00 $\pm$ 0.00	44.00 $\pm$ 0.20	41.00 $\pm$ 0.20	49.03
Gemini 2.0 Flash [37]	53.27 $\pm$ 0.12	60.33 $\pm$ 0.12	74.00 $\pm$ 0.20	76.47 $\pm$ 0.50	46.33 $\pm$ 0.42	56.20 $\pm$ 0.20	58.27 $\pm$ 0.46	53.80 $\pm$ 0.20	59.80
Gemini 1.5 Pro [11]	53.73 $\pm$ 0.46	62.40 $\pm$ 1.31	69.60 $\pm$ 1.31	72.27 $\pm$ 0.83	44.07 $\pm$ 1.01	52.60 $\pm$ 1.00	58.27 $\pm$ 2.14	54.00 $\pm$ 1.00	58.37
Claude 3.5 Sonnet [1]	40.33 $\pm$ 0.42	50.6 $\pm$ 0.20	59.13 $\pm$ 0.90	62.00 $\pm$ 0.53	44.13 $\pm$ 1.03	49.4 $\pm$ 1.25	50.73 $\pm$ 3.19	41.20 $\pm$ 0.60	49.69
Open-Source									
InternVL3-14B[55]	44.73 $\pm$ 1.50	54.93 $\pm$ 1.42	68.13 $\pm$ 0.81	73.73 $\pm$ 0.99	35.60 $\pm$ 1.11	53.00 $\pm$ 0.20	45.67 $\pm$ 0.58	48.33 $\pm$ 0.99	53.02
Qwen2.5-VL-7B [3]	50.87 $\pm$ 0.23	53.33 $\pm$ 0.23	69.60 $\pm$ 0.20	75.93 $\pm$ 0.46	35.93 $\pm$ 0.12	47.87 $\pm$ 0.31	46.07 $\pm$ 0.12	41.27 $\pm$ 0.23	52.61
Qwen2-VL-7B [39]	53.67 $\pm$ 0.90	56.07 $\pm$ 1.72	67.13 $\pm$ 0.76	67.47 $\pm$ 1.14	32.87 $\pm$ 0.99	38.07 $\pm$ 2.42	43.27 $\pm$ 1.72	42.27 $\pm$ 1.01	50.10
LLaVA-OneVision-7B [20]	39.87 $\pm$ 0.12	50.73 $\pm$ 0.64	67.60 $\pm$ 0.69	68.80 $\pm$ 0.35	34.87 $\pm$ 0.64	40.87 $\pm$ 0.31	49.20 $\pm$ 0.20	42.93 $\pm$ 0.70	49.36
InternVL2-8B[38]	42.20 $\pm$ 0.53	44.20 $\pm$ 0.53	61.67 $\pm$ 0.31	64.67 $\pm$ 0.23	33.40 $\pm$ 0.53	38.53 $\pm$ 0.76	43.67 $\pm$ 0.50	41.13 $\pm$ 0.31	46.18
LLaVA-NeXT-7B [21]	34.67 $\pm$ 0.92	33.87 $\pm$ 0.46	57.33 $\pm$ 0.12	62.27 $\pm$ 0.23	30.27 $\pm$ 0.23	39.07 $\pm$ 0.46	47.20 $\pm$ 0.35	40.67 $\pm$ 0.58	43.17
Mantis-8B-Idetics2 [15]	28.07 $\pm$ 0.70	35.47 $\pm$ 0.23	53.73 $\pm$ 0.12	56.53 $\pm$ 0.42	35.67 $\pm$ 0.31	37.73 $\pm$ 0.64	41.53 $\pm$ 0.23	32.53 $\pm$ 0.90	40.16
Deepseek-VL-Chat-7B [27]	32.60 $\pm$ 0.72	34.27 $\pm$ 0.42	51.80 $\pm$ 0.53	52.47 $\pm$ 0.23	32.80 $\pm$ 0.87	29.80 $\pm$ 0.53	41.00 $\pm$ 0.40	36.60 $\pm$ 1.22	38.92
Qwen-VL-Chat-7B [2]	25.20 $\pm$ 1.04	26.60 $\pm$ 0.92	33.60 $\pm$ 1.11	36.80 $\pm$ 1.22	21.07 $\pm$ 2.69	21.73 $\pm$ 1.03	29.47 $\pm$ 1.70	30.53 $\pm$ 0.76	28.13

Table 1 reports model accuracy across categories, each consisting of 500 egocentric (Ego) and 500 exocentric (Exo) questions. Even the best-performing model, GPT-4o, achieves only 60.90% accuracy on E3VQA, underscoring the benchmark’s difficulty. Among open-source models, InternVL3-14B attains the highest accuracy, while Qwen2.5-VL-7B delivers competitive performance despite its smaller number of parameters. Overall, LVLMs struggle the most with numerical reasoning yet perform relatively well on object and attribute recognition. Notably, models consistently underperform on egocentric questions compared to exocentric questions, highlighting difficulties in resolving the first-person perspective.

4.2 Performance Evaluation of M3CoT

To demonstrate the effectiveness of our M3CoT prompting scheme, we compare our technique with three recent multimodal CoT techniques (DDCoT, CoCoT, and CCoT) on the E3VQA benchmark. Experiments are conducted using two leading LVLMs, GPT-4o and Gemini 2.0 Flash, both of which achieve the best performance on E3VQA. Table 2 presents category-wise accuracy for egocentric and exocentric questions. M3CoT improves over CCoT by 4.84 % on GPT-4o and 5.94 % on Gemini 2.0 Flash. In addition, it surpasses DDCoT and CoCoT by 4.15% and 5.71% on GPT-4o, and by 5.03% and 5.81% on Gemini, respectively. The substantial gains in Numerical Reasoning (6.88% and 8.13% over CCoT) highlight M3CoT’s ability to integrate multi-view information for a more complete and accurate understanding. In contrast to existing methods, which exhibit limited or inconsistent improvements and occasionally even show performance drops, M3CoT achieves consistent and substantial gains across all categories. These results validate the effectiveness of our approach in addressing the limitations of current multimodal CoT techniques. For additional results on open-source LVLMs, please refer to Appendix C.2.

5 Analysis

5.1 Analysis of Automated QA Generation Pipeline

To examine how the source of distractors affects the question difficulty, we sample 160 questions (40 per category) and construct four alternative option sets. In each set, all four answer choices are drawn from a single source: text-only, ego view, exo view, or both views. This setup contrasts with our standard configuration, where each distractor is drawn from a different source. As shown in Figure 5(a), the model’s error rate increases in the following order: text-only, both-view, single-view, and our composite setting. This result highlights that constructing answer options from diverse

Table 2: Performance comparison of recent multimodal CoT methods on top-performing models.

Methods	Pose & Action		Object & Attribute		Numerical		Spatial		Avg.
	Ego	Exo	Ego	Exo	Ego	Exo	Ego	Exo	
GPT-4o									
Default	49.87 $\pm$ 1.10	63.47 $\pm$ 0.81	72.47 $\pm$ 0.12	77.00 $\pm$ 0.40	48.47 $\pm$ 1.14	57.67 $\pm$ 0.12	61.13 $\pm$ 0.99	57.13 $\pm$ 0.12	60.90
DDCoT [54]	55.20 $\pm$ 0.72	69.53 $\pm$ 0.31	73.80 $\pm$ 0.92	78.80 $\pm$ 0.40	48.13 $\pm$ 0.64	57.87 $\pm$ 0.58	67.27 $\pm$ 0.76	64.87 $\pm$ 0.50	64.43
CoCoT [50]	50.93 $\pm$ 0.31	66.80 $\pm$ 0.69	72.20 $\pm$ 0.69	76.33 $\pm$ 0.64	49.93 $\pm$ 1.75	60.93 $\pm$ 0.90	65.07 $\pm$ 1.10	60.80 $\pm$ 0.40	62.87
CCoT [32]	55.53 $\pm$ 0.81	67.47 $\pm$ 0.31	73.00 $\pm$ 0.69	77.67 $\pm$ 0.61	48.27 $\pm$ 1.86	62.27 $\pm$ 0.12	63.73 $\pm$ 0.90	62.00 $\pm$ 0.53	63.74
M3CoT (Ours)	58.40 $\pm$ 0.28	69.40 $\pm$ 1.13	78.90 $\pm$ 0.42	82.80 $\pm$ 1.70	56.40 $\pm$ 1.13	67.90 $\pm$ 0.71	71.90 $\pm$ 2.69	62.90 $\pm$ 0.14	68.58
Gemini 2.0 Flash									
Default	53.27 $\pm$ 0.12	60.33 $\pm$ 0.12	74.00 $\pm$ 0.20	76.47 $\pm$ 0.50	46.33 $\pm$ 0.42	56.20 $\pm$ 0.20	58.27 $\pm$ 0.46	53.80 $\pm$ 0.20	59.80
DDCoT [54]	55.60 $\pm$ 0.72	62.60 $\pm$ 1.04	75.53 $\pm$ 0.95	81.13 $\pm$ 0.95	46.13 $\pm$ 0.95	54.67 $\pm$ 1.29	57.47 $\pm$ 0.42	55.60 $\pm$ 1.22	61.09
CoCoT [50]	55.40 $\pm$ 0.40	61.67 $\pm$ 0.50	73.80 $\pm$ 1.25	77.93 $\pm$ 0.12	45.27 $\pm$ 1.21	56.20 $\pm$ 0.35	58.67 $\pm$ 0.42	53.53 $\pm$ 1.01	60.31
CCoT [32]	55.93 $\pm$ 0.46	61.53 $\pm$ 0.31	71.47 $\pm$ 0.76	76.93 $\pm$ 0.58	46.67 $\pm$ 0.70	60.73 $\pm$ 1.10	57.27 $\pm$ 2.27	50.93 $\pm$ 0.61	60.18
M3CoT (Ours)	57.80 $\pm$ 0.20	65.80 $\pm$ 0.20	78.80 $\pm$ 0.72	82.80 $\pm$ 0.40	55.60 $\pm$ 1.06	67.40 $\pm$ 1.25	62.67 $\pm$ 0.81	58.07 $\pm$ 1.10	66.12

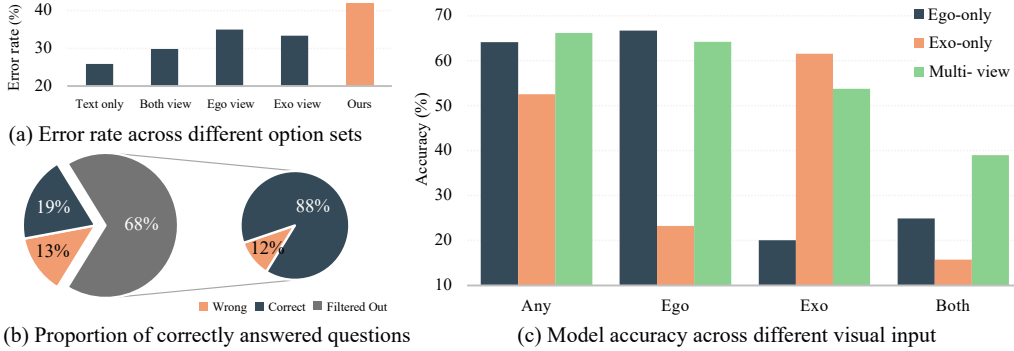


Figure 5: Analysis of the benchmark construction pipeline and model performance under varied input conditions: (a) error rate across option-generation strategies, (b) proportion of correctly answered questions between retained and excluded questions, and (c) performance across different visual input modalities.

input sources increases question difficulty, thereby providing a more rigorous evaluation of LVLMs’ multi-view reasoning abilities.

To assess the effectiveness of our question filtering process, we compare model accuracy on two subsets: a filtered subset (68% of the data) removed during our filtering process and an unfiltered subset (the remaining 32%). As shown in Figure 5(b), 42% of questions in the unfiltered subset are answered incorrectly with ego-exo multi-view input, whereas only 12% of questions in the filtered subset are answered incorrectly. This substantial gap indicates that the filtering pipeline effectively removes questions solvable by superficial cues while retaining more challenging ones that better evaluate a model’s reasoning ability. As a result, the subsequent human verification process becomes more reliable and efficient.

5.2 Analysis of LVLM Performance under Single and Multi-View Inputs

We partition the E3VQA benchmark into four subsets, Any, Ego, Exo, and Both, based on which view(s) are required to answer each question. Specifically, questions in the Any subset can be answered using either the egocentric or exocentric image alone, as each view individually contains sufficient information. The Ego and Exo subsets, in turn, require only the egocentric or exocentric image, respectively, since all relevant information appears in a single view. The Both subset requires integrating cues from both egocentric and exocentric views. We then investigate how the model’s responses differ between single-view and multi-view inputs within each subset. As shown in Figure 5(c), in the Any subset, providing both images yields a marginal accuracy gain, indicating that

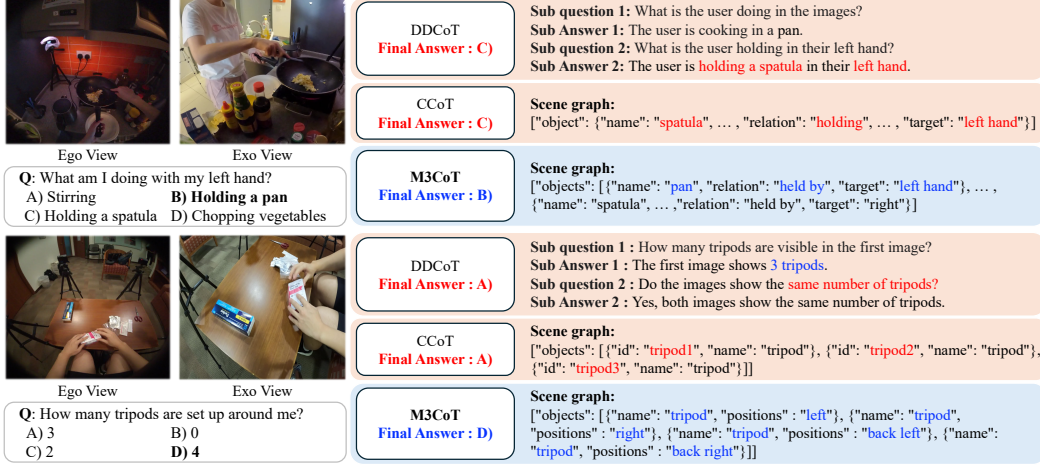


Figure 6: Qualitative examples of answers and reasoning processes generated by different prompting methods. Blue/Red words indicate the key cues that lead to correct/wrong answers.

consistent cues across views can help reinforce the model’s prediction of the correct answer. In the Ego and Exo subsets, providing both images leads to a noticeable performance drop, suggesting that redundant context may confuse the model. In the Both subset, multi-view inputs improve accuracy compared to single-view setups; however, performance remains low, staying below 40%.

5.3 Analysis of Multi-Perspective Scene Graph Generation Strategies

To analyze the advantages of our three scene graph generation perspectives (Ego&Exo, Ego2Exo, and Exo2Ego), we report in Table 3 their respective performances across the Any, Ego, Exo, and Both subsets. The Ego&Exo strategy achieves the largest accuracy gain in the Both subset, demonstrating its strength in integrating complementary cues across viewpoints. In contrast, Ego2Exo performs best on the Exo subset, while Exo2Ego yields the highest improvement on the Ego subset, reflecting their specialization in inferring one view from the other. These results highlight that different scene graph generation strategies provide complementary advantages depending on where the key information required to answer the question is located within the ego-exo view images. In addition, our scene graph refinement stage improves performance beyond any individual strategy by combining their strengths and compensating for their limitations. Overall, these findings confirm that fusing diverse scene graph perspectives produces more robust reasoning in ego-exo multi-view scenarios.

5.4 Qualitative Examples

Figure 6 shows qualitative examples of answers and reasoning processes generated by DDCoT, CCoT, and M3CoT methods using Gemini 2.0 Flash. Although CCoT produces plausible scene graphs, it fails to integrate information across multiple views, resulting in incorrect answers. Specifically, it often misidentifies the same object observed from different perspectives as separate entities. In contrast, our method effectively extracts key information, aligns observations across views, and accurately identifies the same object to answer the question through a multi-perspective, multi-turn reasoning process.

Table 3: Performance comparison of M3CoT’s three perspectives across subsets grouped by the image view(s) required to answer.

Perspective	Required View(s)				Avg.
	Any	Ego	Exo	Both	
Default	66.29	64.83	54.21	37.49	59.80
Ego&Exo	66.29	67.44	56.67	<u>50.87</u>	<u>63.65</u>
Ego2Exo	<u>68.08</u>	62.71	<u>61.51</u>	43.91	62.83
Exo2Ego	66.94	<u>68.02</u>	59.92	39.13	62.98
M3CoT (ours)	69.79	69.28	62.91	53.04	66.12

6 Related Work

6.1 Ego-Exo Datasets and Tasks

Egocentric and exocentric views offer complementary information for understanding users and their environments. Early datasets like Charades-Ego [35] and LEMMA [14] introduced paired ego-exo data, while Ego-Exo4D [12] further scaled this paired ego-exo data with large, synchronized videos capturing diverse real-world scenarios. To generalize semantic understanding across multiple perspectives, a body of work has focused on learning view-invariant representations [40, 45]. Furthermore, efforts to align ego-exo content have emerged, including object-level mappings [10] and techniques for identifying and segmenting camera wearers in exocentric scenes [9, 53]. In parallel, cross-view knowledge transfer has been actively explored, with each perspective leveraged to improve the understanding of the other [51, 22, 43, 34]. Several studies have addressed viewpoint selection across perspectives by proposing methods for dynamically selecting informative views over time [29, 30]. Others have explored generating egocentric videos from exocentric video inputs using diffusion-based models [26, 44] or cropping exocentric image frames to distill egocentric-relevant cues [6]. Despite these advances, a task that jointly reasons over synchronized egocentric and exocentric views within LVLMs remains underexplored, highlighting a promising direction for future research.

6.2 Visual Question Answering with LVLMs

Visual Question Answering (VQA) benchmarks test LVLMs’ ability to interpret and reason over diverse visual content. Most existing VQA benchmarks are constructed from large-scale web-crawled data, typically consisting of images captured from fixed third-person cameras [42, 48, 49]. To support scenarios that require understanding from the user’s perspective, egocentric VQA benchmarks capturing first-person views have been introduced. EgoVQA [8] evaluates first-person visual understanding capabilities by offering both egocentric and exocentric queries on first-person visual inputs. EgoSchema [31] evaluates long-form egocentric video understanding by assessing a model’s ability to recall previously observed objects and events. EgoThink [4] evaluates first-person reasoning capabilities across diverse categories that reflect practical real-world scenarios. Another line of work includes embodied QA benchmarks such as EmbodiedQA [5] and OpenEQA [28], where agents are required to navigate or interact with their environments to answer given queries. Although numerous VQA benchmarks aim to evaluate LVLMs across diverse aspects, no existing benchmark assesses a model’s ability to seamlessly combine complementary visual information from paired ego and exo views. For a detailed comparison of VQA benchmarks, please refer to Table 4 in the Appendix.

6.3 Chain-of-Thought Prompting in LVLMs

Building on its success in large language models (LLMs), Chain-of-Thought (CoT) prompting has been extended to LVLMs to enhance inference-time reasoning. DDCoT breaks down a question into a sequence of sub-questions and corresponding sub-answers, which are then used collectively to derive the final answer to the original question [54]. CoCoT, introduced for multi-image input scenarios, compares the similarities and differences between images, guiding the model to answer questions based on the identified visual contrasts [50]. CCoT facilitates understanding the overall context of an image through scene graphs, which are first generated by the LVLM and then incorporated into the prompt to enable compositional reasoning over objects, relations, and attributes [32]. Despite their successes, their applicability to ego-exo multi-image contexts remains unexplored, raising an open challenge for extending CoT methods to these settings.

7 Conclusion

In this work, we introduced E3VQA, the first benchmark that systematically assesses whether LVLMs can reason jointly over egocentric and exocentric views. By curating 4K high-quality question-answer pairs grounded in synchronized ego-exo images, E3VQA serves as a rigorous testbed for multi-view understanding. In addition, we proposed M3CoT, a novel prompting strategy that merges scene graphs from diverse perspectives into a unified graph. Extensive experiments demonstrate that M3CoT consistently outperforms the strong CCoT baseline, underscoring the importance of multi-perspective integration for multi-view understanding. By establishing a benchmark for evaluating LVLMs’ ego-exo reasoning and enhancing their multi-view understanding ability, this work takes a step toward more context-aware visual assistants capable of operating in complex, real-world environments.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) (RS-2024-00398157) and IITP under the artificial intelligence semiconductor support program to nurture the best talents (IITP-2023-RS-2023-00256081), and by the National Research Foundation of Korea (NRF) (RS-2023-00208985), all funded by the Korea government (MSIT).

References

- [1] Anthropic. Claude 3.5 sonnet model card addendum, 2024.
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.
- [4] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302, 2024.
- [5] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018.
- [6] Zi-Yi Dou, Xitong Yang, Tushar Nagarajan, Huiyu Wang, Jing Huang, Nanyun Peng, Kris Kitani, and Fu-Jen Chu. Unlocking exocentric video-language data for egocentric video representation learning. *arXiv preprint arXiv:2408.03567*, 2024.
- [7] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332, 2023.
- [8] Chenyou Fan. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.
- [9] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5125–5133, 2017.
- [10] Yuqian Fu, Runze Wang, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. Objectrelator: Enabling cross-view object relation understanding in ego-centric and exo-centric videos. *arXiv preprint arXiv:2411.19083*, 2024.
- [11] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [12] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.
- [13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [14] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *European Conference on Computer Vision*, pages 767–786. Springer, 2020.
- [15] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024.
- [16] Donghoon Kim, Minji Bae, Kyuhong Shim, and Byonghyo Shim. Visually guided decoding: Gradient-free hard prompt inversion with language models. *arXiv preprint arXiv:2505.08622*, 2025.

- [17] Donghoon Kim, Gusang Lee, Kyuhong Shim, and Byonghyo Shim. Preserving pre-trained representation space: On effectiveness of prefix-tuning for large multi-modal models. *arXiv preprint arXiv:2411.00029*, 2024.
- [18] Jiseob Kim, Kyuhong Shim, Junhan Kim, and Byonghyo Shim. Vision transformer-based feature extraction for generalized zero-shot learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [19] Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*, 2022.
- [20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [21] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.
- [22] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021.
- [23] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.
- [24] Chao Liu, Chi San Cheung, Mingqing Xu, Zhongyue Zhang, Mingyang Su, and Mingming Fan. Toward facilitating search in vr with the assistance of vision large language models. In *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology*, pages 1–14, 2024.
- [25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [26] Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Zheng Shou. Exocentric-to-egocentric video generation. *Advances in Neural Information Processing Systems*, 37:136149–136172, 2024.
- [27] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- [28] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498, 2024.
- [29] Sagnik Majumder, Tushar Nagarajan, Ziad Al-Halah, and Kristen Grauman. Switch-a-view: Few-shot view selection learned from edited videos. *arXiv preprint arXiv:2412.18386*, 2024.
- [30] Sagnik Majumder, Tushar Nagarajan, Ziad Al-Halah, Reina Pradhan, and Kristen Grauman. Which viewpoint shows it best? language for weakly supervising view selection in multi-view videos. *arXiv preprint arXiv:2411.08753*, 2024.
- [31] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.
- [32] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024.
- [33] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023.

- [34] Dominick Reilly, Manish Kumar Govind, Le Xue, and Srijan Das. From my view to yours: Ego-augmented learning in large vision language models for understanding exocentric daily living activities. *arXiv preprint arXiv:2501.05711*, 2025.
- [35] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.
- [36] Alessandro Suglia, Claudio Greco, Katie Baker, Jose L. Part, Ioannis Papaioannou, Arash Eshghi, Ioannis Konstantas, and Oliver Lemon. AlanaVLM: A multimodal embodied AI foundation model for egocentric video understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11101–11122. Association for Computational Linguistics, November 2024.
- [37] Gemini Team. Gemini: A family of highly capable multimodal models, 2024.
- [38] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024.
- [39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [40] Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3307–3317, 2023.
- [41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [42] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.
- [43] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024.
- [44] Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. *arXiv preprint arXiv:2504.11732*, 2025.
- [45] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710, 2023.
- [46] Jingkan Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European conference on computer vision*, pages 178–196. Springer, 2022.
- [47] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022.
- [48] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041, 2021.
- [49] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.
- [50] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*, 2024.
- [51] Haoyu Zhang, Qiaohui Chu, Meng Liu, Yunxiao Wang, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, Yaowei Wang, and Liqiang Nie. Exo2ego: Exocentric knowledge guided mllm for egocentric video understanding. *arXiv preprint arXiv:2503.09143*, 2025.

- [52] Juexiao Zhang, Gao Zhu, Sihang Li, Xinhao Liu, Haorui Song, Xinran Tang, and Chen Feng. Multiview scene graph. *Advances in Neural Information Processing Systems*, 37:17761–17788, 2024.
- [53] Ziwei Zhao, Yuchen Wang, and Chuhua Wang. Fusing personal and environmental cues for identification and segmentation of first-person camera wearers in third-person views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16477–16487, 2024.
- [54] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibe Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.
- [55] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingdong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction explicitly state that LVLMs struggle to integrate egocentric and exocentric views, introduce the E3VQA benchmark with 4K high-quality QA pairs, and propose the M3CoT prompting method which yields significant performance gains, fully matching the theoretical framework and empirical results presented later in the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper includes a “Limitations” section in Appendix that discusses the reliance of the E3VQA benchmark on the EgoExo4D dataset—highlighting potential biases in recording environments and action diversity—and outlines future work to mitigate this dependency and broaden dataset coverage.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper is proposing the E3VQA benchmark and its dataset creation pipeline, as well as introducing the M3CoT prompting method—and does not involve any formal theoretical results or proofs, so this question is not applicable.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides full disclosure of the main experimental results with error bars, clearly identifies each LVM variant evaluated, and includes all system and user prompt templates along with detailed specifications in the Appendix, enabling faithful reproduction of the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All data including E3VQA and code are publicly accessible, and all related materials—including the prompt templates—can be found in Appendix.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We evaluate off-the-shelf LVLMS using their default settings (no additional training) and provide full details of our dataset construction and test setup: all system and user prompt templates are documented in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We include error bars ($\pm$) alongside all main accuracy results—capturing variability across test samples or repeated runs—in Table 1 and Table 2, enabling readers to assess statistical significance of the reported improvements

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide detailed specifications of our compute environment in Appendix, ensuring all reported results can be faithfully reproduced

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We fully adhere to the NeurIPS Code of Ethics in all aspects—including data collection and usage, privacy safeguards, and licensing—and provide discussion of these ethical considerations in Appendix.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We outline both potential benefits and risks associated with our E3VQA benchmark and M3CoT method in the Ethics Statement in Appendix, covering improved scene understanding applications alongside considerations of data privacy, bias, and potential misuse.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The proposed E3VQA dataset is based on the existing dataset EgoExo4D dataset, and our paper does not introduce any new pretrained language models or image generators. Therefore, our research is free from the risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: Reference 11 cites EgoExo4D, and Section 3 explicitly states that the dataset is based on it.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: Section 3 provides details on the generation pipeline of the proposed E3VQA dataset.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: E3VQA datasets are created from existing EgoExo4D dataset, so crowdsourcing was not used.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper uses LLMs for dataset generation and multi-image VQA, as described in Sections 3 and 5.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Table 4: Comparison between E3VQA and existing VQA benchmarks. Unlike other benchmarks, E3VQA is designed to evaluate comprehensive scene understanding and reasoning across diverse question perspectives using paired ego-exo images.

Benchmark	Task Objective	Visual Perspective	Question Perspective	Answer Type	Evaluator	#Questions (test)
MSVD-QA [42]	General Understanding	Exo	Exo	Predefined-Label	Accuracy	13K
MSRVTT-QA [42]	General Understanding	Exo	Exo	Predefined-Label	Accuracy	72K
Social-IQ [49]	Social Understanding	Exo	Exo	Multi-Choice	Accuracy	7.5K
Pano-AVQA [48]	Spatial / Audio-Visual Reasoning	Exo	Exo	Predefined-Label	Accuracy	5.3K
EgoVQA [8]	Egocentric Visual Understanding	Ego	Ego or Exo	Multi-Choice	Accuracy	120
EgoSchema [31]	Long-Term Reasoning	Ego	Ego	Multi-Choice	Accuracy	5K
EgoThink [4]	First-Person Thinking	Ego	Ego	Open-Ended	LLMs	700
EmbodiedQA [5]	Goal-Driven Scene Understanding	Ego	Exo	Predefined-Label	Accuracy	529
OpenEQA [28]	Environment Understanding	Ego	Ego or Exo	Open-Ended	LLMs	1.6K
E3VQA	Comprehensive Scene Understanding and Reasoning	Ego and Exo	Ego or Exo	Multi-Choice	Accuracy	4K

A Additional Details of the E3VQA Benchmark

A.1 Categories and Challenges

In addition to the challenges described in Section 2.2, each of the following four categories highlights a distinct challenge in the ego-exo multi-image scenario:

- **Pose & Action Perception** focuses on recognizing a person’s physical state and movement, such as how their body is positioned and what kinds of gestures or actions they are performing. The presence of multiple people, including the user and duplicated individuals across views, can confuse the model when identifying the question’s target. The model must correctly identify the intended individuals and interpret their physical state and behavior.
- **Object & Attribute Perception** involves identifying objects and their attributes, such as color, pattern, or type. Objects may appear in only one view, be partially occluded, or look different due to variations in viewpoint and field of view. To answer correctly, models must resolve such ambiguities and ground the object consistently across views.
- **Numerical Reasoning** addresses tasks involving counting and comparing quantities, such as determining the number of people or objects in a scene. A single view may not include all instances necessary to answer the question, and the same object may appear redundantly across different views. To produce accurate counts, the model must integrate information from both views by handling overlapping objects and aggregating evidence across views.
- **Spatial Reasoning** focuses on understanding the spatial information of a scene, including how objects and people are positioned relative to one another and how they are arranged within the environment. In multi-view spatial reasoning, differences in viewpoint angle and field of view can cause the same object to appear at varying positions in each image, become occluded in some views, or exhibit different spatial relationships with surrounding objects. To overcome these challenges, the model must align positional information from multiple views to construct a coherent understanding of spatial relationships within the scene.

A.2 Dataset Composition and Statistics

Figure 7 illustrates the statistics of the E3VQA benchmark. Figure 7(a) shows the distribution of correct answer positions among the four options (A-D) for each category. The uniform distribution of correct answers indicates that the dataset is balanced with respect to answer positions. Figure 7(b) shows the distribution of source video types used to construct E3VQA, demonstrating the benchmark’s broad coverage of real-world user-interaction scenarios. Finally, Figure 7(c) illustrates the detailed composition of question types within each category, underscoring E3VQA’s broad scope of evaluation.

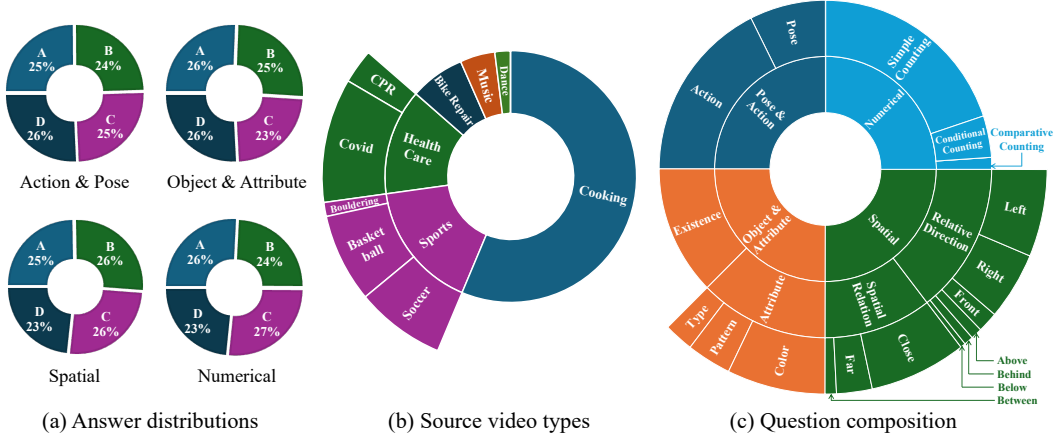


Figure 7: E3VQA statistics: (a) Distribution of the correct answers among the four options (A-D), (b) Distribution of source video types used to construct E3VQA, and (c) Composition of question types within each category.

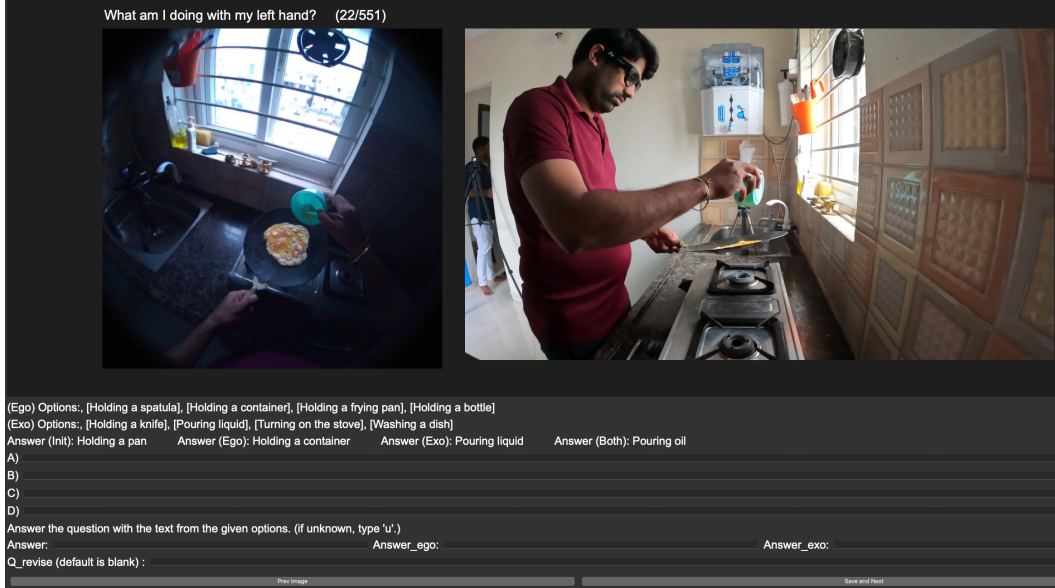


Figure 8: User interface used by annotators during human verification.

A.3 Details of Human Verification

The human verification is conducted by four members of our research team (the authors), all of whom have domain expertise in vision-language reasoning and extensive experience with ego-exo data. Each annotator is assigned to a specific question category and conducts verification using the user interface shown in Figure 8. Annotators first remove questions that are ambiguous or unanswerable when both ego and exo views are provided. Next, they discard overly simple questions that can be solved through shallow pattern matching, such as cases where A_{both} is correct regardless of reasoning, even if they pass the question filtering stage. Then, annotators utilize the view-specific responses generated in Section 2.3.2 to construct answer options. To supplement potentially redundant or low-quality responses, they additionally use two option sets, each containing four candidate options derived from the ego and exo images, respectively. Using these option sets, the annotators curate complete question sets that include the question, the correct answer, and the distractor options. Finally, each annotator

Table 5: Comparison of open-source LVLMs in terms of architecture (vision encoder and LLM) and use of egocentric data during training.

Model	Vision Encoder	LLM Backbone	Train w/ Ego Data
InternVL3-14B	InternViT-300M-448px-V2.5	Qwen2.5-14B	Not Provided
Qwen2.5-VL-7B	ViT (customized)	Qwen2.5-7B	Not Provided
Qwen2-VL-7B	ViT-L	Qwen2-7B	✗
LLaVA-NeXT-OneVision-7B	SigLIP-SO	Qwen2-7B	✓
InternVL2-8B	InternViT-300M	Qwen2.5-7B	✓
LLaVA-NeXT-Interleave-7B	SigLIP-SO	Qwen1.5-7B	✗
MANTIS-Idefics2-8B	SigLIP	Mistral-7B-v0.1	✗
Deepseek-VL-chat-7B	SigLIP-L, SAM-B	DeepSeek-LLM-7B	✗
Qwen-VL-Chat-7B	ViT-bigG	Qwen-7B	✗

Table 6: Performance comparison of recent closed and open-source models on E3VQA in the text-only setting (without visual input).

LVLMs	Action		Object		Numerical		Spatial		Avg.
	Ego	Exo	Ego	Exo	Ego	Exo	Ego	Exo	
Closed-Source									
Gemini 2.0 Flash	20.20	24.80	20.60	22.80	27.80	30.20	16.20	15.20	22.23
GPT-4o	25.00	30.40	12.20	23.20	7.40	30.80	11.60	15.40	19.50
Open-Source									
InternVL3-14B	31.00	37.20	33.80	30.20	38.40	32.80	10.80	11.80	28.25
Qwen2.5-VL-7B	32.20	34.60	32.80	31.00	35.20	36.20	8.00	5.40	26.93
LLaVA-OneVision-7B	32.60	38.00	31.40	26.80	27.40	28.80	7.00	5.60	24.70

reviews subsets created by others to ensure consistency and clarity across the dataset, filtering out any ambiguous or low-quality instances.

A.4 Robustness to Temporal Misalignment

To evaluate the robustness of the E3VQA benchmark under temporal misalignment between egocentric and exocentric views, we analyze how synchronization gaps affect the consistency of ground-truth answers. Specifically, we randomly sample 80 examples (10 from each category) and evaluate how the ground truth changes when the ego and exo views are misaligned by 1 to 3 seconds. When the synchronization gap is 1 or 2 seconds, 99% and 97% of the ground-truth answers remain unchanged, respectively. Even with a 3-second gap, 94% of the answers remain consistent. Although maintaining perfect temporal alignment can be challenging in real-world applications, our analysis shows that small temporal mismatches (e.g., a few seconds) have minimal impact on the consistency of the ground truth.

B Experimental Details

B.1 LVLMs and Evaluation Setup

Table 5 provides an overview of the LVLMs used in our experiments in terms of the model architecture and the use of egocentric data in training. These models are selected based on their capability of processing multi-image inputs. By evaluating models with diverse vision-language architectures, we examine how recent LVLMs respond to and reason through the challenges posed by the E3VQA benchmark. For evaluation, we use NVIDIA RTX A6000 GPUs. All evaluation results are reported as the mean and standard deviation over three independent runs, using each model’s default generation settings.

B.2 Text-Only Baseline Results

To investigate potential biases in the E3VQA benchmark, we evaluate recent LVLMs under a text-only setting, where only the textual question is provided without any visual input. This experiment allows us to examine whether models can exploit linguistic patterns or biases present in the data without

Table 7: Performance comparison of recent multimodal CoT methods on E3VQA constructed from LEMMA.

Methods	Pose & Action		Object & Attribute		Numerical		Spatial		Avg.
	Ego	Exo	Ego	Exo	Ego	Exo	Ego	Exo	
GPT-4o									
Default	39.68	43.55	63.49	72.58	56.45	46.03	39.68	45.16	50.83
DDCoT [54]	38.10	48.39	66.67	70.97	64.52	47.62	47.62	38.71	52.83
CoCoT [50]	44.44	46.77	69.84	70.97	54.84	47.62	46.03	51.61	54.02
CCoT [32]	36.51	48.39	68.25	69.35	53.23	47.62	47.62	40.32	51.41
M3CoT (Ours)	44.44	51.61	71.43	75.81	59.68	65.08	63.49	50.00	60.19
Gemini 2.0 Flash									
Default	39.68	53.23	66.67	74.19	51.61	46.03	46.03	46.77	53.03
DDCoT [54]	39.68	46.77	60.32	77.42	54.84	47.62	55.56	40.32	52.82
CoCoT [50]	46.03	48.39	65.08	74.19	50.00	47.62	47.62	40.32	52.41
CCoT [32]	44.44	51.61	66.67	72.58	56.45	49.21	46.03	40.32	53.41
M3CoT (Ours)	44.44	59.68	66.67	72.58	56.45	42.86	55.56	50.00	56.03

Table 8: Performance comparison of recent multimodal CoT methods on open-source models.

Methods	Pose & Action		Object & Attribute		Numerical		Spatial		Avg.
	Ego	Exo	Ego	Exo	Ego	Exo	Ego	Exo	
InternVL3 - 14B									
Default	44.73 $\pm$ 1.50	54.93 $\pm$ 1.42	68.13 $\pm$ 0.81	73.73 $\pm$ 0.99	35.60 $\pm$ 1.11	53.00 $\pm$ 0.20	45.67 $\pm$ 0.58	48.33 $\pm$ 0.99	53.02
DDCoT [54]	47.87 $\pm$ 0.83	58.33 $\pm$ 2.64	68.47 $\pm$ 0.50	72.67 $\pm$ 1.42	35.33 $\pm$ 2.53	46.80 $\pm$ 2.12	50.67 $\pm$ 1.10	45.93 $\pm$ 0.95	53.26
CoCoT [50]	49.53 $\pm$ 0.81	57.27 $\pm$ 0.50	68.27 $\pm$ 1.14	72.53 $\pm$ 1.14	34.87 $\pm$ 1.55	47.93 $\pm$ 0.64	49.20 $\pm$ 1.91	46.27 $\pm$ 0.95	53.23
CCoT [32]	44.60 $\pm$ 1.91	58.40 $\pm$ 2.09	65.27 $\pm$ 0.64	73.80 $\pm$ 0.53	37.80 $\pm$ 3.30	50.00 $\pm$ 0.72	46.27 $\pm$ 1.33	48.80 $\pm$ 1.78	53.12
M3CoT (Ours)	45.87 $\pm$ 1.21	60.00 $\pm$ 0.35	70.60 $\pm$ 0.40	75.73 $\pm$ 0.76	35.07 $\pm$ 0.50	50.87 $\pm$ 0.70	50.80 $\pm$ 0.92	49.40 $\pm$ 0.72	54.79
InternVL3 - 8B									
Default	43.70 $\pm$ 3.25	54.90 $\pm$ 0.42	64.80 $\pm$ 0.85	70.30 $\pm$ 0.71	35.90 $\pm$ 2.12	45.20 $\pm$ 1.13	42.10 $\pm$ 2.97	46.60 $\pm$ 3.11	50.44
DDCoT [54]	48.10 $\pm$ 0.99	59.20 $\pm$ 4.24	67.20 $\pm$ 0.28	68.80 $\pm$ 1.13	34.60 $\pm$ 0.85	47.00 $\pm$ 0.00	46.30 $\pm$ 2.69	45.80 $\pm$ 1.41	52.13
CoCoT [50]	43.90 $\pm$ 0.14	58.20 $\pm$ 1.13	65.10 $\pm$ 0.42	68.40 $\pm$ 1.13	37.10 $\pm$ 1.84	48.70 $\pm$ 0.42	43.50 $\pm$ 2.97	43.40 $\pm$ 0.57	51.04
CCoT [32]	44.00 $\pm$ 0.85	55.00 $\pm$ 1.41	63.60 $\pm$ 0.57	68.30 $\pm$ 1.27	35.50 $\pm$ 0.42	51.20 $\pm$ 1.41	43.40 $\pm$ 0.57	44.70 $\pm$ 3.54	50.71
M3CoT (Ours)	45.50 $\pm$ 0.14	57.20 $\pm$ 0.00	68.20 $\pm$ 0.00	71.20 $\pm$ 0.00	37.60 $\pm$ 0.00	47.50 $\pm$ 0.71	53.80 $\pm$ 0.00	49.20 $\pm$ 0.00	53.78

relying on visual reasoning. The results in Table 6 show that all models exhibit substantially lower performance in the absence of visual information, confirming that the E3VQA benchmark requires ego-exo image-grounded reasoning rather than relying solely on textual cues.

C Additional Experiments and Analysis of M3CoT

C.1 Generalization across Datasets

To validate the generalization of our proposed framework across different datasets, we further extend E3VQA using LEMMA [14], a multi-view dataset that captures goal-directed daily activities in home environments. Unlike Ego-Exo4D, the ego images in LEMMA are in a standard rectangular format (i.e., rectified) without the dark peripheral regions. Following the same data generation pipeline in Section 2.3, we construct 500 additional samples evenly distributed across all categories.

As shown in Table 7, the results demonstrate that our E3VQA benchmark construction pipeline generalizes well across datasets, indicating that the difficulty of ego-exo multi-view reasoning arises not from dataset-specific visual properties but from the inherent challenge of integrating complementary perspectives for visual understanding. Furthermore, the consistent improvements achieved by M3CoT across datasets highlight the robustness of its multi-perspective reasoning mechanism.

Table 9: Performance comparison across question subsets grouped by required views.

LVLMS	Methods	Any	Ego	Exo	Both	Avg.
GPT-4o	Default	65.20	63.96	58.59	39.28	56.76
	DDCoT	69.71	69.37	60.83	37.68	59.40
	CoCoT	66.94	64.41	61.47	43.48	59.08
	CCoT	67.94	67.50	61.05	42.03	59.63
	M3CoT	73.45	74.30	64.23	52.61	66.15
Gemini 2.0 Flash	Default	65.81	66.08	54.38	37.52	55.95
	DDCoT	66.48	66.36	58.22	31.13	55.55
	CoCoT	67.23	65.43	56.44	32.43	55.38
	CCoT	65.76	66.18	55.29	43.21	57.61
	M3CoT	69.76	69.24	62.97	53.19	63.79

C.2 Evaluation on Open-Source Models

We present experimental results of our M3CoT prompting technique compared to existing CoT methods on open-source LVLMS. Specifically, we apply M3CoT on InternVL3-14B [55], the top-performing open-source model, and further evaluate the performance on InternVL3-8B. As shown in Table 8, most CoT methods result in only marginal performance gains, with several failing to improve accuracy and even causing degradation in certain categories. This aligns with prior findings suggesting that the CoT method is often ineffective in smaller models with limited reasoning capability [41, 19, 7]. Despite the limitations observed in smaller models, our M3CoT consistently achieves superior performance compared to other CoT methods, highlighting its robustness across model sizes.

C.3 Performance across Question Subsets Based on Required Views

We further compare model performance across the four subsets (Any, Ego, Exo, and Both) introduced in Section 5.2 to examine how our M3CoT approach provides advantages across different question types. Among these, performance on the Both subset reflects a model’s ability to integrate multi-view information. Results on the Any subset reflect how well a model handles redundant or overlapping information, while those on the Ego and Exo subsets indicate a model’s robustness when unnecessary or irrelevant information is included in the input. As shown in Table 9, M3CoT consistently outperforms other CoT-based methods across all subsets, demonstrating its overall effectiveness in multi-view reasoning.

C.4 Evaluation of Scene Graph Quality

The quality of scene graphs is often evaluated using the intersection-over-union (IoU) with ground-truth scene graphs [46, 52]. However, ground-truth scene graphs often include overly exhaustive details that may not always be beneficial for answering questions accurately. To better assess the practical utility of scene graphs generated by M3CoT, we introduce two complementary metrics:

- **False Discovery Rate (FDR)** is measured as the proportion of incorrect elements among the predicted scene graph components, including object classes, relations, and attributes. This metric measures how accurately the scene graphs represent the visual scene without including non-existent elements.
- **Answer Accuracy** is measured by model accuracy when the generated scene graph serves as input to the model, indicating how effectively the graph supports reasoning when solving questions.

To evaluate the scene graphs with respect to these two metrics, we constructed a subset of 120 samples from E3VQA, evenly distributed across all question categories. We then evaluated the scene graphs before and after the refinement step of M3CoT. After refinement, the FDR of the generated scene graphs decreased from 9.37% to 6.21%, while the answer accuracy increased from 46.88% to 60.42%. These results demonstrate that the refinement stage of M3CoT not only reduces errors in the scene graphs but also reorganizes their visual information to better support the model in generating accurate answers.

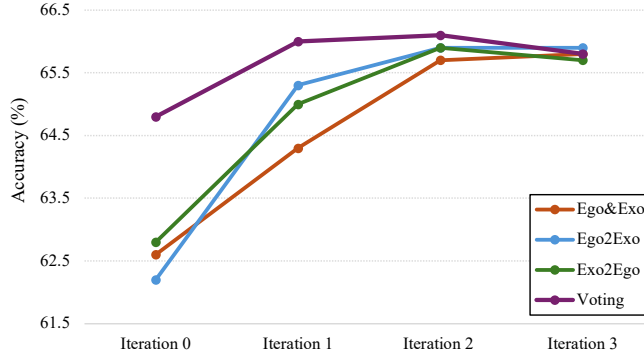


Figure 9: Performance across different perspectives and majority voting results over iteration steps.

Table 10: Comparison of computational overhead between M3CoT and existing CoT prompting methods on GPT-4o and Gemini 2.0 Flash.

LVLMS	Method	Token Usage	Time (s)	Acc. (%)
GPT-4o	Default	2,120.25	2.62	60.90
	DDCoT	6,807.18	9.57	64.43
	CoCoT	2,349.23	6.25	62.87
	CCoT	9,225.30	22.59	63.74
	M3CoT	47,250.12	37.56	68.58
Gemini 2.0 Flash	Default	790.60	2.56	59.80
	DDCoT	2,752.09	5.48	60.31
	CoCoT	950.71	3.84	61.09
	CCoT	2,817.97	7.14	60.18
	M3CoT	26,164.36	20.03	66.12

C.5 Analysis of Iteration Steps in Multi-Agent Scene Graph Refinement

To analyze the effect of iteration steps in M3CoT, we report the accuracy of each individual perspective as well as the majority-voted answer derived from them at each iteration. As shown in Figure 9, without any information exchange across perspectives, all individual perspectives and the majority-voted answer achieve relatively low accuracy (iteration 0). As agents begin to exchange their scene graphs, we observe a steady improvement in the accuracy of each individual perspective, suggesting that iterative refinement facilitates mutual enhancement through shared contextual understanding. This process also leads to a corresponding increase in voting accuracy, reflecting not only the enhanced quality of individual predictions but also a stronger consensus across perspectives. However, beyond the second iteration, we find that both individual accuracy and voting accuracy plateau. We attribute this saturation to the convergence of information across agents: while initial iterations benefit from the diversity of complementary perspectives, excessive alignment diminishes the gains from their integration. This observation highlights a trade-off in our multi-perspective refinement strategy between refining individual scene representations and preserving representational diversity. Note that all experiments and analyses in this paper are conducted with a fixed iteration count of 1, using Gemini 2.0 Flash unless otherwise specified.

C.6 Analysis of Computational Cost

To examine the trade-off between computational efficiency and performance, we measure inference time, token usage, and accuracy of recent CoT prompting methods, including M3CoT, on GPT-4o and Gemini 2.0 Flash (see Table 10). Please note that the reported inference time may be unreliable due to variability in API response, so it should be interpreted as a general trend rather than an exact value. While multi-perspective reasoning requires additional computation, it enables a more accurate understanding than other recent prompting methods.

D Qualitative Examples of M3CoT

D.1 Comparison with Other CoT Methods

We provide additional examples that illustrate how M3CoT improves reasoning compared to other CoT approaches (see Figure 10).

D.2 Scene Graphs from Three Perspectives

To further examine how different perspectives in M3CoT contribute to capturing complementary information, we present additional qualitative examples of scene graphs derived from each perspective. As shown in Figures 11, 12, and 13, the scene graphs from the three perspectives exhibit complementary strengths depending on the question, particularly regarding which image should be referenced to answer it.

D.3 Conflict Resolution during Refinement

We provide qualitative examples illustrating how M3CoT resolves conflicts among scene graphs from different perspectives during the refinement stage (see Figure 14).

D.4 Typical Failure Cases

We provide qualitative examples illustrating the failure cases of M3CoT (see Figure 15).

E Prompt Templates

E.1 Prompt Templates for E3VQA Construction

To guide LVLMS in understanding the question categories and tasks for generating meaningful question-answer pairs, we carefully design the prompts for each stage. To generate question-answer pairs from a single viewpoint, we use the prompts shown in Figure 16–25. For view-specific response generation, we apply the prompts in Figure 26–33. For response-based filtering, we use the prompts shown in Figure 34 and 35. Finally, to generate four candidate options from either the ego or exo image, we use the prompts illustrated in Figure 36 and 37.

E.2 Prompt Templates for Baselines and CoT Methods

The system and user prompts used in the baseline experiments of E3VQA are shown in Figure 38. In addition, for M3CoT, the prompts for scene graph generation from each perspective are shown in Figure 39–41, and the prompts for scene graph refinement across agents are presented in Figure 42. The prompts used in other CoT baselines are shown in Figure 43–45.

F Limitations

Despite its contributions, this work has several limitations. First, the E3VQA benchmark is constructed solely from the Ego-Exo4D and LEMMA datasets, which may exhibit dataset bias and limited generalizability in diverse real-world scenarios. Second, although the queries and answer options in E3VQA are carefully crafted, they may not fully capture the diversity of natural language expressions and user intents encountered in real-world interactions with visual AI assistants. Third, while recent AI APIs offer a solution for scaling the benchmark, their use entails substantial financial costs. Fourth, M3CoT introduces increased computational overhead due to its multi-step reasoning across multiple perspectives, which may limit its applicability in resource-constrained scenarios. Finally, since E3VQA is constructed from images rather than videos, the benchmark may not fully assess an LVLMS’s ability to capture temporal cues and motion dynamics, an aspect we leave for future work.

G Ethics Statement

This work has the potential to positively impact society by enhancing the capabilities of visual assistants and embodied AI systems, particularly in scenarios that require comprehensive scene understanding from both egocentric and exocentric views. Such advancements may enhance human-AI interaction and improve support in assistive technologies. However, the use of egocentric visual data may raise important privacy concerns, especially in sensitive environments. We acknowledge these risks and emphasize the importance of implementing safeguards and transparency mechanisms in future deployments. As part of our commitment to responsible data use, we have obtained the appropriate licenses from the contributing institutions for the use of the Ego-Exo4D dataset in this research.

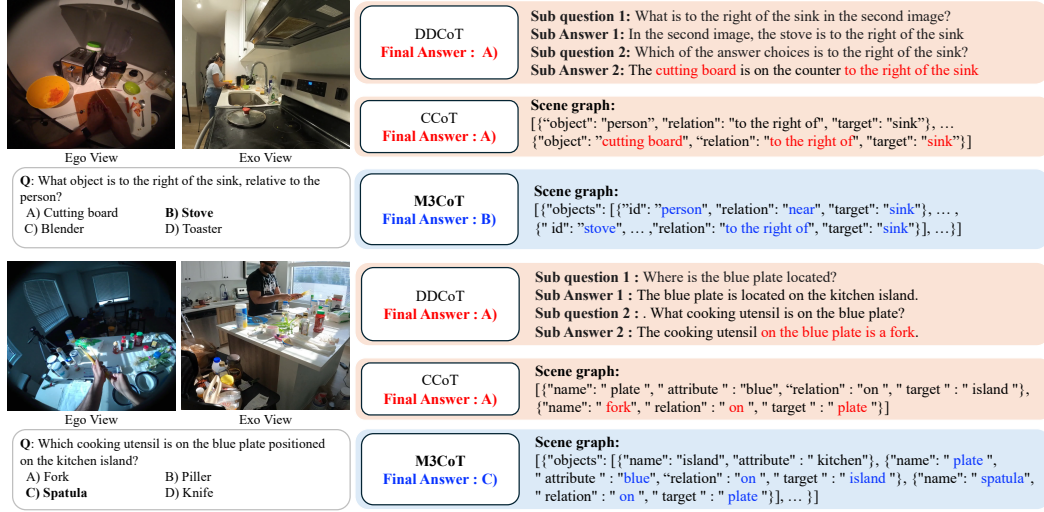


Figure 10: Qualitative examples of answers and reasoning processes generated by different prompting methods.

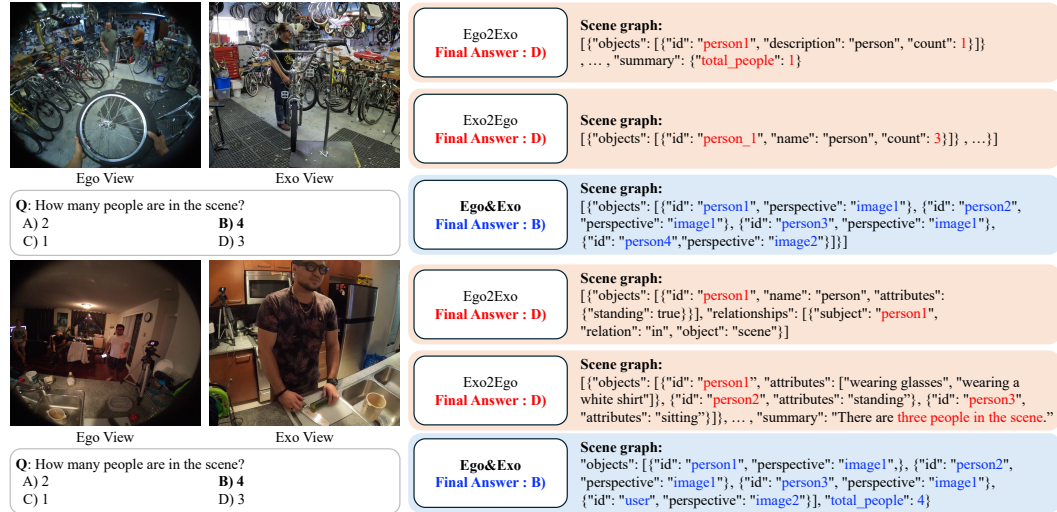


Figure 11: Qualitative examples of answers and reasoning processes generated by different perspectives. The scene graph from the Ego&Exo perspective demonstrates a strong capability to capture the information necessary for answering questions grounded in both ego and exo views.

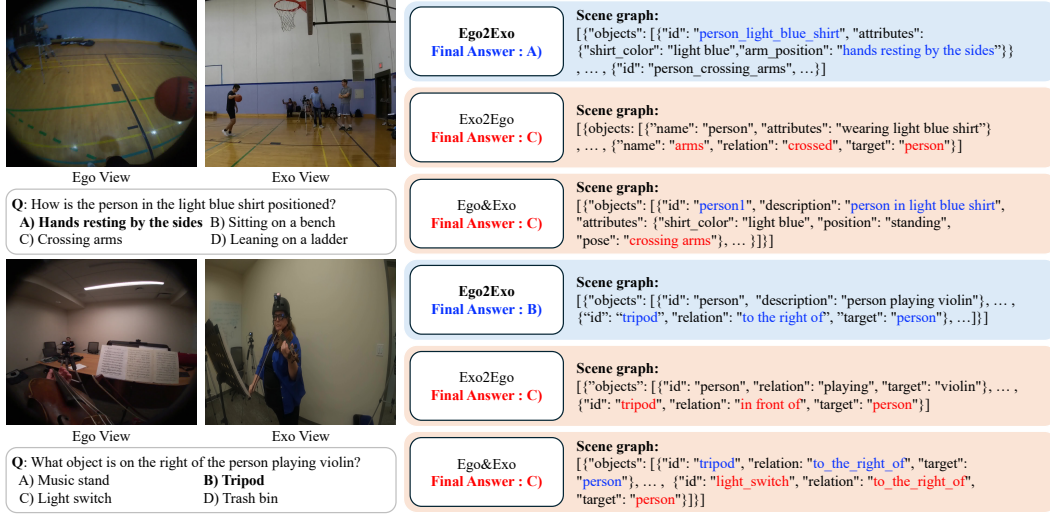
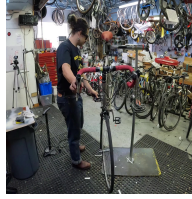


Figure 12: Qualitative examples of answers and reasoning processes generated by different perspectives. The scene graph from the Ego2Exo perspective demonstrates a strong capability to capture the information necessary for answering questions grounded in the exo view alone.



Figure 13: Qualitative examples of answers and reasoning processes generated by different perspectives. The scene graph from the Exo2Exo perspective demonstrates a strong capability to capture the information necessary for answering questions grounded in the ego view alone.

		Ego&Exo Answer : D)	Scene graph: [{"objects": [{"id": "bicycle", "description": "bicycle being worked on", "attributes": {"wheels_attached": 1}}]}]
Ego View	Exo View		
Q: How many wheels are attached to the bicycle? A) 0 C) 3 B) 2 D) 1		Ego&Exo Final Answer: B)	Scene graph: [{"objects": [{"object": "wheel_1", "relation": "attached to", "target": "bicycle"}, {"object": "wheel_2", "relation": "attached to", "target": "bicycle"}, ...]}] ↓ Scene graph: [{"objects": [{"id": "wheel_1", "description": "rear wheel of the bicycle"}, {"id": "wheel_2", "description": "front wheel of the bicycle"}]}]
		Exo2Ego Final Answer: B)	Scene graph: [{"objects": [{"id": "wheel_1", "relation": "attached to", "target": "bicycle"}, {"id": "wheel_2", "relation": "attached to", "target": "bicycle"}, ...]}]

		Ego2Exo Answer : B)	Scene graph: [{"objects": [{"id": "person", "description": "wearing gray shirt", "relationships": [{"predicate": "doing", "object": "washing dishes"}]}]}]
Ego View	Exo View		
Q: What is the person in the gray shirt doing with their hands? A) Cleaning the countertop C) Stirring with a spoon B) Washing dishes D) Scrubbing a dish		Ego2Exo Final Answer: C)	Scene graph: [{"objects": [{"id": "person", "description": "wearing gray shirt", "relationships": [{"predicate": "doing", "object": "stirring with a spoon"}]}]}] ↓ Scene graph: [{"objects": [... , {"object": "hands", "relation": "doing", "target": "stirring with a spoon"}, ...]}]
		Ego2Exo Final Answer: C)	Scene graph: [{"objects": [{"id": "person", "description": "wearing gray shirt", "relationships": [{"predicate": "doing", "object": "stirring with a spoon"}]}]}]
		Exo2Ego Final Answer: C)	Scene graph: [{"objects": [{"object": "person", "attributes": "wearing gray shirt", "relationships": [{"predicate": "doing", "object": "stirring with a spoon"}]}]}]

Figure 14: Qualitative examples illustrating how conflicts among scene graphs from different perspectives are resolved by M3CoT. Scene graphs with inaccurate elements are refined during M3CoT’s refinement stage, yielding accurate and consistent representations.

		Ego2Exo Final Answer : C)	Scene graph: [{"objects": [... , {"object": "bowl", "attribute": "yellow"}, ... , {"object": "person", "relation": "using", "target": "knife"}]}]
Ego View	Exo View	Exo2Ego Final Answer : C)	Scene graph: [{"objects": [{"object": "person", "relation": "using", "target": "knife"}, ... , {"object": "small bowl", "attributes": "green"}]}]
Q: What item is the person using to cook? A) Fork B) Spatula C) Knife D) Spoon		Ego&Exo Final Answer : C)	Scene graph: [{"objects": [{"id": "person", "description": "A person cooking", "relationship": [{"type": "USING", "target": "knife"}]}]}]
		Ego2Exo Final Answer : A)	Scene graph: [{"objects": [{"object": "cutting board", "relation": "to the left of", "target": "I"}, {"object": "trash can", "relation": "to the left of", "target": "I"}]}]
Ego View	Exo View	Exo2Ego Final Answer : B)	Scene graph: [{"objects": [{"id": "cutting_board", ... , "location": "right of user"}, {"id": "trash_can", "perspective": "image2", "location": "to the left from the user"}]}]
Q: What object is to my left? A) Cutting board B) Trash can C) Recipe paper D) Paper towel roll		Ego&Exo Final Answer : A)	Scene graph: [{"relationships": [{"subject": "cutting_board", "relation": "to the right of", "object": "user"}, {"subject": "trash_can", "relation": "to the left of", "object": "user"}]}]
		Ego2Exo Answer : B)	Scene graph: [{"objects": [{"id": "hand", "attributes": ["right"], "relationships": [{"predicate": "reaching", "object": "surface"}]}]}]
Ego View	Exo View	Ego&Exo Answer: A)	Scene graph: [{"relationships": [... , {"source": "right_hand", "target": "spray_paint", "type": "holding"}, ...]}]
Q: What am I doing with my right hand? A) Spraying B) Reaching C) Cleaning D) Pointing		Ego2Exo Final Answer: A)	Scene graph: [{"relationships": [... , {"source": "right_hand", "target": "spray_paint", "type": "near"}, ...]}]
		Ego&Exo Final Answer: A)	Scene graph: [{"relationships": [... , {"source": "right_hand", "target": "spray_paint", "type": "holding"}, ...]}]

Figure 15: Qualitative examples illustrating M3CoT’s failure cases. The top row shows examples where all three perspectives produce incorrect predictions. The middle row illustrates cases where the scene graphs are correctly generated but fail to yield accurate answers. The bottom row shows cases where the refinement stage fails due to erroneous scene graphs.

Egocentric Single-View QA Generation Prompt
{Ego Image} You are given the visual input from the camera worn by the user (referred to as 'I'). Based on this visual input, generate three question-answer pairs. Ensure that the generated question-answer pairs are directly based on the visual input. {Category-wise Prompt} Requirements: Each question must explicitly include the pronoun 'I' or 'me' to ensure the focus remains on the user. Each answer should be a single word or a short phrase. Ensure that all three question-answer pairs meet these criteria and are relevant to the visual input. Strictly adhere to the format of the provided examples.

Figure 16: Egocentric single-view QA generation prompt.

Egocentric Single-View QA Generation Prompt: Action & Pose
Instructions : Each question must focus on my actions , body posture , or gestures . The answer must be a verb or verb phrase (e.g., writing , stretching , crossing arms). Do not generate QA pairs with overly generic answers like ‘standing’ or ‘reaching’. Question Categories & Templates: Actions (What am I doing?) – What am I doing? – What am I doing with my [body part]? Body Posture (How am I positioned?) – How is my body positioned? – How am I sitting /standing /lying? – What is my posture? Gestures (What movement am I making?) – What am I doing with my hands? – What gesture am I making? – How am I moving my arms/legs/head? Examples: Q: How is my body positioned? A: Sitting cross-legged Q: What am I doing with my left hand? A: Holding a book Q: What gesture am I making? A: Waving

Figure 17: Egocentric single-view QA generation prompt: Pose & Action.

Egocentric Single-View QA Generation Prompt: Object & Attribute
Instructions : Each question must focus on identifying a specific object (e.g., mug cup, laptop) or describing an attribute of an object (e.g., navy blue , striped pattern) associated with me. The answer must be a noun or noun phrase , avoiding overly generic responses such as something or object. Question Categories & Templates: Object Identification (What am I interacting with?) - What am I holding? - What object is on the table beside me? - Which item am I picking up? Object Attributes (What does it look like?) - What color is the shirt I am wearing? - What pattern is on my jacket? - What type of shoes am I wearing? Examples: Q: What color is the shirt I am wearing? A: Navy blue Q: Which object am I holding in my right hand? A: A small notebook Q: What pattern does my sweater have? A: Checkered pattern

Figure 18: Egocentric single-view QA generation prompt: Object & Attribute.

Egocentric Single-View QA Generation Prompt: Spatial

Instructions :

Each question must focus on the spatial relationships between me and objects in my surroundings.

The answer must be a specific object or location descriptor (e.g., coffee cup, bookshelf, under the table).

Do not generate QA pairs with overly generic answers.

Question Categories & Templates:

Object Proximity (What is closest or farthest?):

- What object is closest to me?
- Which object is the farthest from me?
- What is the nearest object to my [body part]?

Relative Positioning (Where are objects located?)

- What object is to my left/right/front/behind?
- Which object is above/below me?
- Spatial Relations (How are objects arranged?)
- Which object is between me and [another object]?

Examples:

Q: What object is closest to my left hand?

A: Coffee cup

Q: Which object is the farthest from me?

A: Bookshelf

Q: What object is on my right side?

A: Tissue

Figure 19: Egocentric single-view QA generation prompt: Spatial.

Egocentric Single-View QA Generation Prompt: Numerical

Instructions :

Each question must focus on numerical reasoning by counting or quantifying specific elements directly related to me. This may include the number of people, objects, or other countable items present in my surroundings. The answer must be a numerical value that accurately represents the count of the indicated elements. Do not generate questions about overly generic objects (e.g., items, objects). All numerical answers must be within the range of 0 to 5.

Question Categories & Templates:

Counting People (How many people are around me?)

- How many people are in the image excluding me?
- How many individuals are facing the same direction as I am?

Counting Objects (How many things are near or with me?)

- How many [objects] am I holding?
- How many [items] are on the table beside me?

Quantitative Comparisons (How do the numbers compare to what I have?)

- How many more books are on my desk than on the shelf?
- By how much does the number of items in my hands exceed the number on the table?

Examples:

Q: How many people are in the image excluding me?
A: 3

Q: How many more bowls are on my table compared to the table behind me?
A: 2

Q: How many apples am I holding?
A: 3

Figure 20: Egocentric single-view QA generation prompt: Numerical.

Exocentric Single-View QA Generation Prompt

{Exo Image}

You are given with the visual input from a fixed-position camera capturing a scene.
Based on this visual input, generate three question-answer pairs.
Ensure that the generated question-answer pairs are directly based on the visual input.

{Category-wise Prompt}

Requirements:
Each answer should be a single word or a short phrase.
Ensure that all three question-answer pairs meet these criteria and are relevant to the visual input.
Strictly adhere to the format of the provided examples.

Figure 21: Exocentric single-view QA generation prompt.

Exocentric Single-View QA Generation Prompt: Action & Pose

Instructions:
Each question must focus on the actions, body posture, or gestures within the scene.
The answer must be a verb or verb phrase (e.g., writing, stretching, crossing arms).
Do not generate QA pairs with overly generic answers like 'standing' or 'reaching'.

Question Categories & Templates:
 Actions (What is the person doing?)
 – What is the [descriptive] person doing?
 – What is the [descriptive] person doing with their [body part]?
 Body Posture (How is the person positioned?)
 – How is the [descriptive] person positioned?
 – What is the posture of the [descriptive] person?
 Gestures (What movements is the person making?)
 – What kind of gesture is the [descriptive] person making?
 – How is the [descriptive] person moving their arms/legs/head?

Examples:
 Q: What is the man sitting in the chair doing?
 A: Watching a phone
 Q: What is the posture of the person wearing a green shirt?
 A: Raising one arm
 Q: What is the woman in the black jacket doing with their right hand?
 A: Holding a book

Figure 22: Exocentric single-view QA generation prompt: Action & Pose.

Exocentric Single-View QA Generation Prompt: Object & Attribute
Instructions : Each question must focus on identifying a specific object in the scene (e.g., 'mug cup', 'laptop') or describing an attribute of an object (e.g., 'navy blue', 'striped pattern'). Questions should reference people or objects by descriptors (e.g., 'the woman in the white top', 'the man with the striped shirt'). The answer must be a noun or noun phrase, avoiding overly generic responses such as 'something' or 'object'. Question Categories & Templates: Object Identification (What is present?) - What is the man with the striped shirt holding? - What object is placed on the table? - Which item is the woman wearing a blue top picking up? Object Attributes (What does it look like?) - What color is the shirt worn by the man wearing a cap? - What pattern is on the jacket worn by the woman carrying a handbag? - What type of shoes is the man standing near the window wearing? Examples: Q: What color is the top worn by the woman holding the towel? A: White Q: Which object is the man in the black shirt holding in his right hand? A: Smartphone Q: What pattern does the sweater worn by the person holding a cup have? A: Checkered pattern

Figure 23: Exocentric single-view QA generation prompt: Object & Attribute.

Exocentric Single-View QA Generation Prompt: Spatial
Instructions: Each question must explicitly reference an object's or a person's spatial relationship within the scene. The answer must be a specific object or location descriptor (e.g., scissors, frying pan, under the table). Do not generate QA pairs with overly generic answers. Question Categories & Templates: Object Proximity (What is closest or farthest?) - Which object is closest to the person wearing [specific item]? - Which object is the farthest from [reference point]? - What is the nearest object to [specific location or object]? Relative Positioning (Where are objects located?) - What object is to the left/right/front/behind of the man with [specific item]? - What object is to the left/right/front/behind [reference object]? - Which object is positioned above/below [reference object]? Spatial Relations (How are objects arranged?) - Which object is positioned between [object A] and [object B]? - What item is placed underneath/inside [object]? - Which object is located between the two people sitting on the \\ bench? Examples: Q: What is the object on the far right of the desk? A: Scissors Q: Which cookware is closest to the woman wearing a striped shirt? A: Frying pan Q: What object is placed directly in front of the man wearing a cap? A: Backpack Q: What object is placed underneath the table? A: Storage box

Figure 24: Exocentric single-view QA generation prompt: Spatial.

Exocentric Single-View QA Generation Prompt: Numerical
Instructions : Each question must focus on numerical reasoning by counting or quantifying specific elements within the scene. This may include the number of people, objects, or other countable items present in the image. The answer must be a numerical value that accurately represents the count of the indicated elements. Do not generate questions about overly generic objects (e.g., items, objects). All numerical answers must be within the range of 0 to 5. Question Categories & Templates: Counting People (How many are there?) – How many people are in the scene? – How many individuals are facing the camera? Counting Objects (How many things are visible?) – How many objects is [person descriptor] holding? – How many items are on the table? Quantitative Comparisons (How do the numbers compare?) – How many more books are on the table than on the shelf? – By how much does the number of items in the man's hands exceed the number on the table? Examples: Q: How many people are in the scene? A: 3 Q: How many objects is the woman in the striped shirt holding? A: 2 Q: How many oranges are placed on the table? A: 5

Figure 25: Exocentric single-view QA generation prompt: Numerical.

View-Specific Response Expansion Prompt: Ego View
{Ego Image} You are given a visual input from a camera worn by the user (referred to as 'I') along with a corresponding question. Based on the visual input, generate the best possible answer. {Category-wise Prompt} Requirements: Each answer option should be a single word or a short phrase. Follow the provided format strictly. Q: {Question}

Figure 26: View-specific response expansion prompt: Ego view.

View-Specific Response Expansion Prompt: Exo View
{Exo Image} You are given a visual input from a fixed-position camera capturing a scene along with a corresponding question. Based on the visual input, generate the best possible answer. {Category-wise Prompt} Requirements: Each answer should be a single word or a short phrase. Follow the provided format strictly. Q: {Question}

Figure 27: View-specific response expansion prompt: Exo view.

View-Specific Response Expansion Prompt: Both Views
{Ego Image} {Exo Image} You are provided with two visual inputs in sequence, each captured from a different perspective: 1. The view from the camera worn by the user ('I'). 2. The view captured by an external camera observing the user ('I'). These two images capture the same event at the same time. Based on the visual inputs, generate the best possible answer. {Category-wise Prompt} Requirements: Each answer should be a single word or a short phrase. Follow the provided format strictly. Q: {Question}

Figure 28: View-specific response expansion prompt: Both views.

View-Specific Response Expansion Prompt: Text Only
Based on the question, generate the best possible answer. {Category-wise Prompt} Requirements: Each answer should be a single word or a short phrase. Follow the provided format strictly. Q: {Question}

Figure 29: View-specific response expansion prompt: text only.

View-Specific Response Expansion Prompt: Action & Pose
Instructions: The answer must be a verb or verb phrase (e.g., writing, stretching, crossing arms). Do not generate overly generic answers like 'standing' or 'reaching'. Output format: Q: How is my body positioned? A: Sitting cross-legged Q: What is the man sitting in the chair doing? A: Watching a phone

Figure 30: View-specific response expansion prompt: Action & Pose.

View-Specific Response Expansion Prompt: Object & Attribute
Instructions: The answer must be a noun or noun phrase, avoiding overly generic responses such as 'something' or 'object'. Output format: Q: What color is the shirt I am wearing? A: Navy blue Q: What color is the top worn by the woman holding the towel? A: White

Figure 31: View-specific response expansion prompt: Object & Attribute.

View-Specific Response Expansion Prompt: Spatial
Instructions: The answer must be a specific object or location descriptor (e.g., coffee cup, bookshelf, under the table). Do not generate overly generic answers. Output format: Q: What object is closest to my left hand? A: Coffee cup Q: What is the object on the far right of the desk? A: Scissors

Figure 32: View-specific response expansion prompt: Spatial.

View-Specific Response Expansion Prompt: Numerical
Instructions: The answer must be a numerical value that accurately represents the count of the indicated elements. All numerical answers must be within the range of 0 to 5. Output format: Q: How many people are in the image excluding me? A: 3 Q: How many people are in the scene? A: 3

Figure 33: View-specific response expansion prompt: Numerical.

Response-Based Question Filtering Prompt 1
Here is the question: '{Question}'. The provided answer is {answer_both}, and the given label is {answer_init}. Do they convey the same meaning based on the question? Respond with a single word or phrase.

Figure 34: Response-based question filtering prompt (1).

Response-Based Question Filtering Prompt 2
Here is the question: '{Question}'. The provided answer is '{answer_text}', and the given label is '{answer_init}'. Do they convey the same meaning based on the question? Respond with a single word or phrase.

Figure 35: Response-based question filtering prompt (2).

Option Generation Prompt: Ego
{Ego Image} You are given a visual input from a camera worn by the user (referred to as 'I'). Based on the following question and answer, generate four multiple-choice options. Question: {Question} Answer: {answer_ego} Ensure that each incorrect option is closely related to the visual content, making it challenging to easily identify the correct answer. Follow the format below exactly: Options: [Option1] [Option2] [Option3] [Option4]

Figure 36: Option generation prompt: Ego.

Option Generation Prompt: Exo
<code>{Exo Image}</code> You are given a visual input from a fixed-position camera capturing a scene. Based on the following question and answer, generate four multiple-choice options. Question: <code>{Question}</code> Answer: <code>{answer_exo}</code> Ensure that each incorrect option is closely related to the visual content, making it challenging to easily identify the correct answer. Follow the format below exactly: Options: [Option1] [Option2] [Option3] [Option4]

Figure 37: Option generation prompt: Exo.

System Prompt & Question (Instruction) Prompt
System Prompt You are a helpful assistant. You are provided with two visual inputs in sequence, each captured from a different perspective: 1. The view from the camera worn by the user ('I'). 2. The view captured by an external camera observing the user ('I'). The first image shows what the user ('I') sees from their perspective . The user's full body cannot be visible; you may only see parts of their body, like their hand, foot, or arm, or in some cases, none of the user's body at all. The second image shows both the user and the environment from a third-person perspective with a broad view. The user's full body is visible, but due to the fixed viewpoint, some parts may not be visible. These two images capture the same event at the same time. Your task is to analyze both images along with the question and provide the most accurate response based on the visual information from both perspectives. Question (Instruction) Prompt {Ego Image} {Exo Image} {Question} Only one option is correct. Present the answer in the form X).

Figure 38: System Prompt and Question(Instruction) Prompt.

M3CoT Prompts - Ego2Exo Perspective

Scene graph generation phase (Ego2Exo)

Task:

For the provided image and its associated question, generate a scene graph in JSON format that includes the following:

1. Objects that are relevant to answering the question.
2. Object attributes that are relevant to answering the question.
3. Object relationships that are relevant to answering the question.

Just generate the scene graph in JSON format. Do not say extra words.

{Ego Image}

{Question Prompt}

Scene graph refinement phase (Ego2Exo)

Task:

For the provided image from a different view and the scene graph generated from the previous view, refine the scene graph in JSON format as follows:

1. Review and Update Existing Objects and Relationships:
Examine the objects and relationships in the initial scene graph. Update their attributes or positions based on observations from both views. Remove only elements that are clearly erroneous (e.g., annotation errors or duplicates).

2. Incorporate New Information:
Identify and add any new objects or relationships that appear in the new view.

3. Align and Reconcile Across Views:
For overlapping objects and relationships, align them using spatial proximity and semantic similarity. If attribute discrepancies arise, select values that best reflect the combined observations.

Ensure that the updated scene graph is logically and physically consistent, avoiding contradictions or impossible configurations. Just generate the refined scene graph in JSON format. Do not say extra words.

{Exo Image}

{Question Prompt}

{Assistant's response(Ego-only SG)}

Initial question response phase (Ego2Exo)

Use the images and the refined scene graph as context and answer the following question.

{Ego Image}

{Exo Image}

{Question Prompt}

{Assistant's response(Refined SG)}

Figure 39: M3CoT prompt (1).

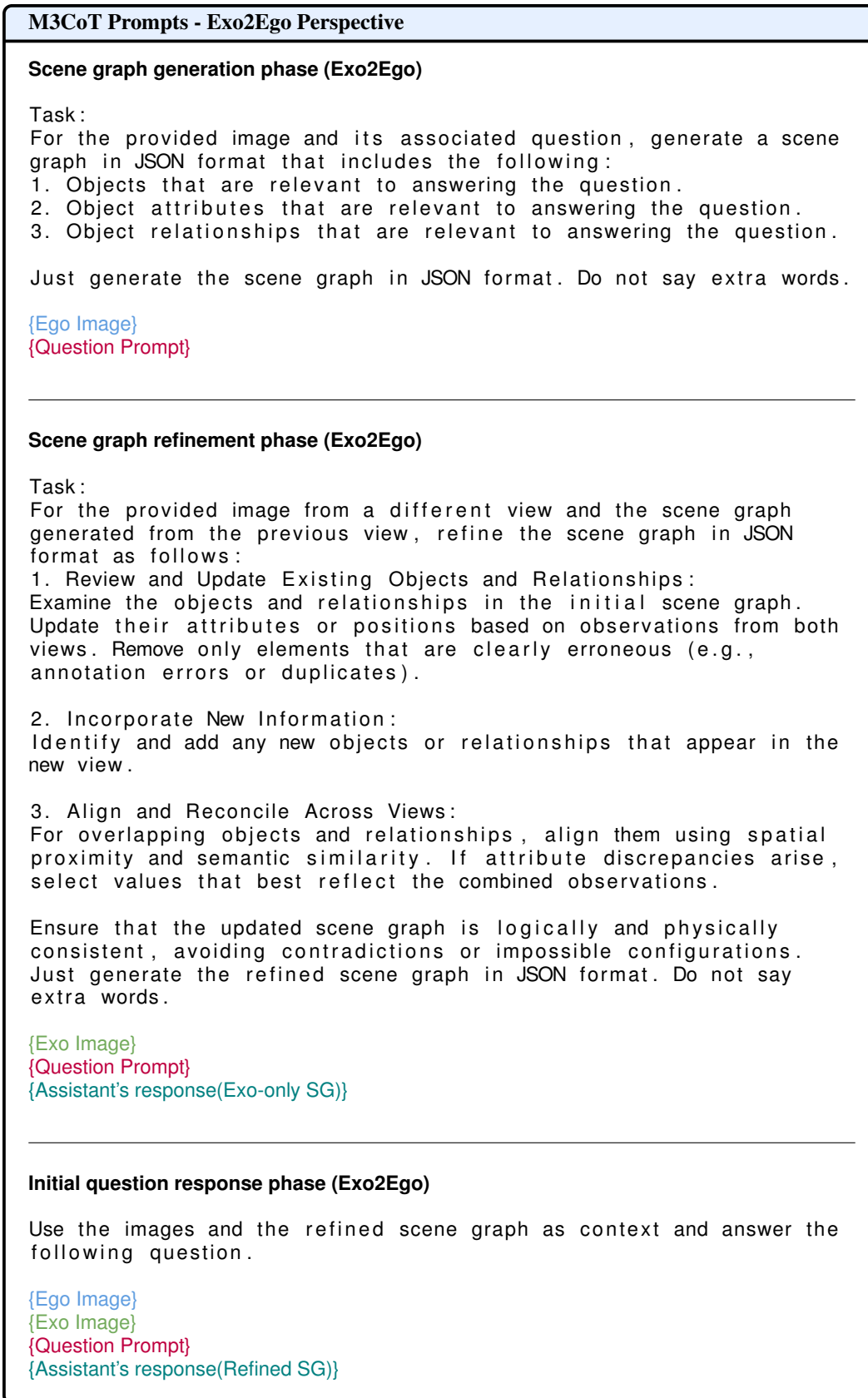


Figure 40: M3CoT prompt (2).

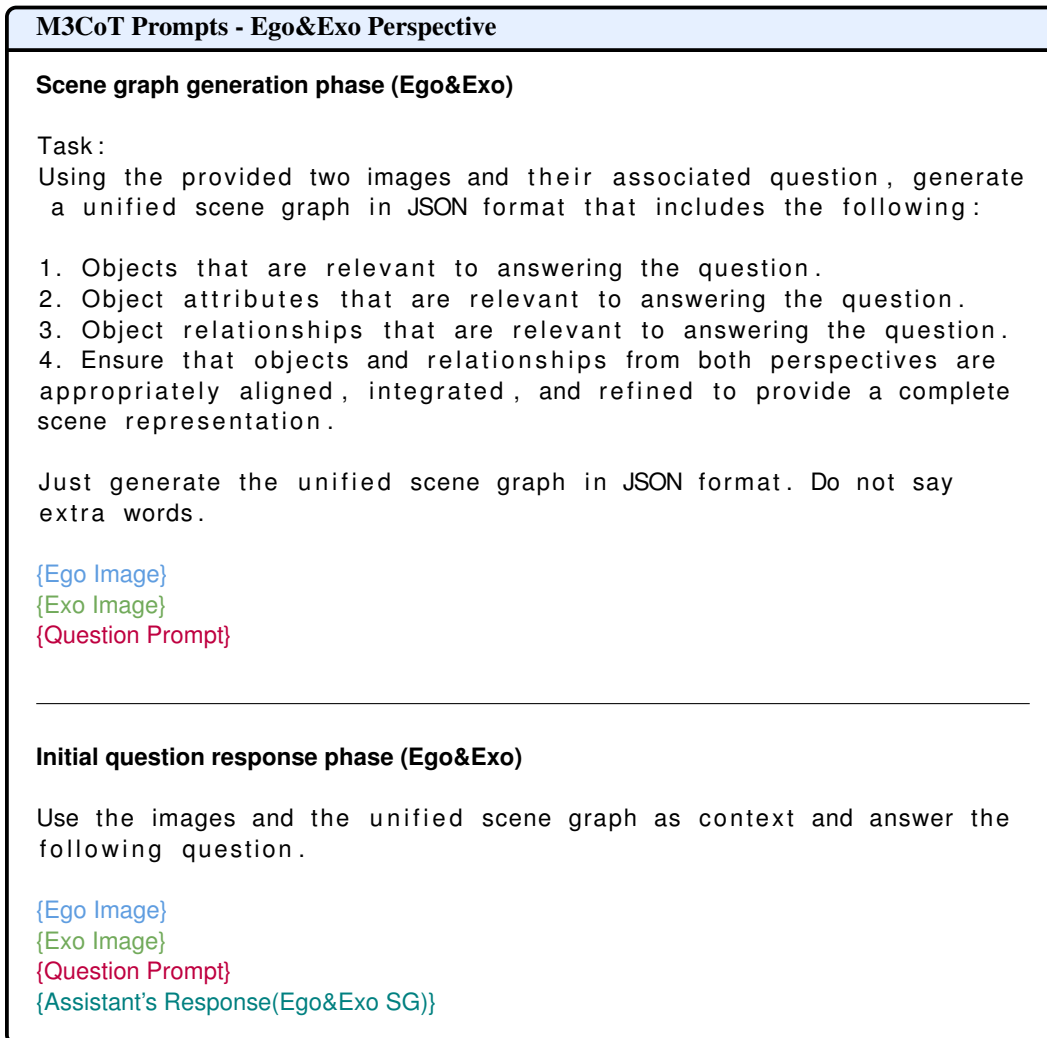


Figure 41: M3CoT prompt (3).

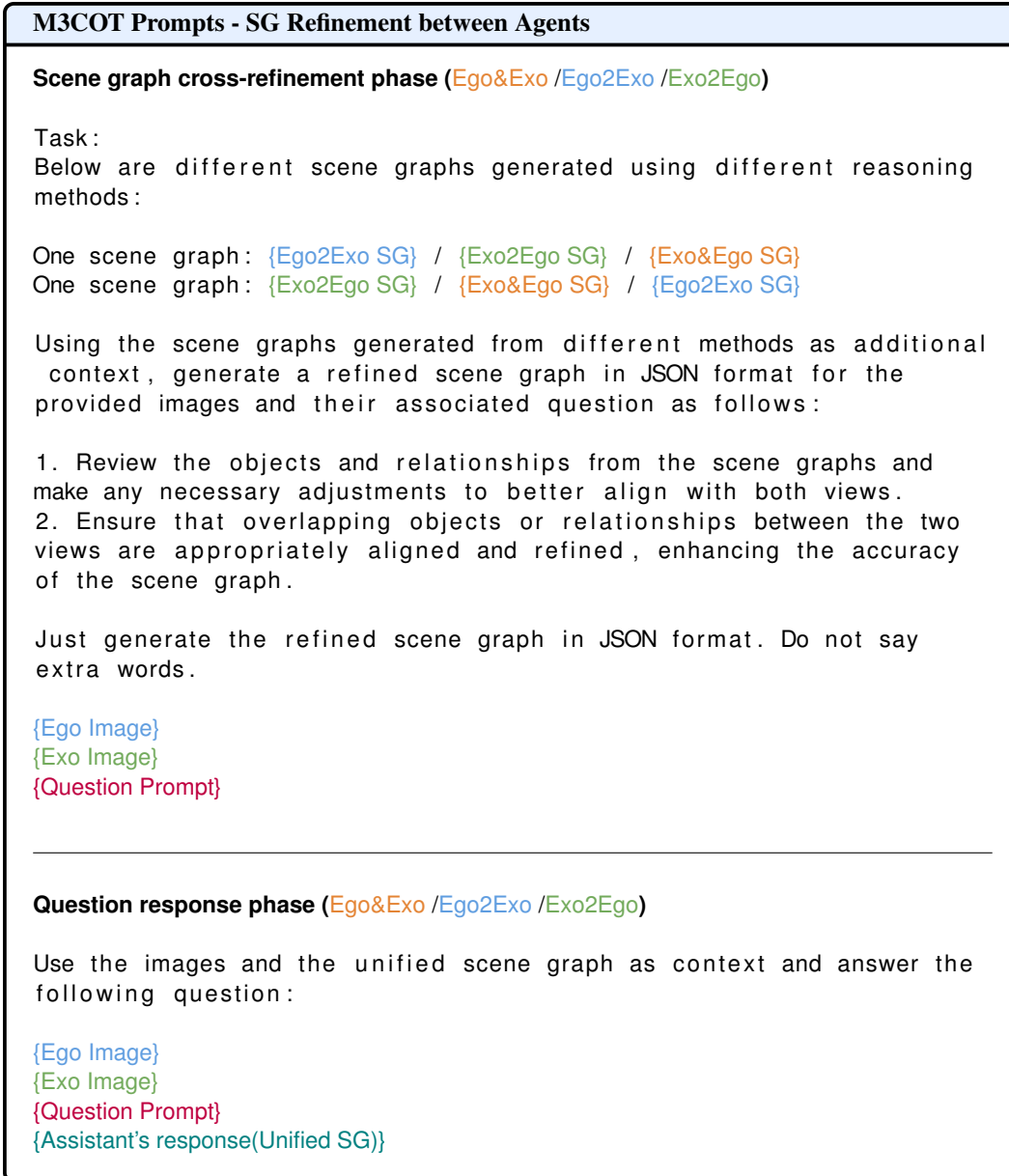


Figure 42: M3CoT prompt (4).

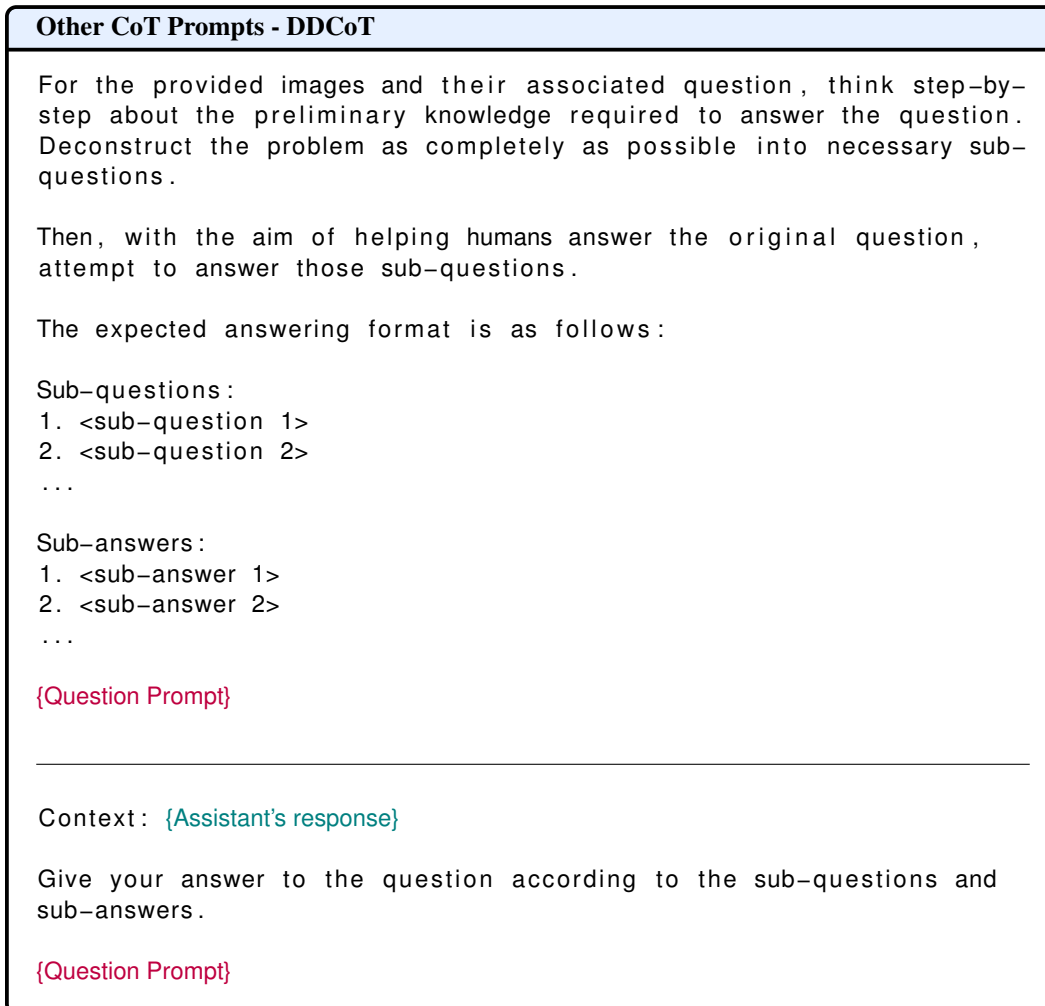


Figure 43: DDCoT Prompt.

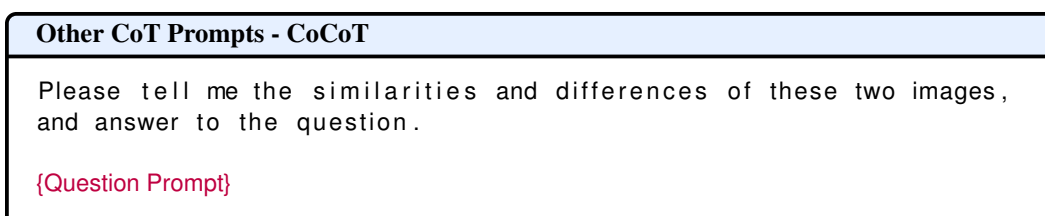


Figure 44: CoCoT Prompt.

Other CoT Prompts - CCoT
For the provided images and their associated question , generate a scene graph in JSON format that includes the following : 1. Objects that are relevant to answering the question . 2. Object attributes that are relevant to answering the question . 3. Object relationships that are relevant to answering the question . Just generate the scene graph in JSON format . Do not say extra words . {Question Prompt} Scene Graph: {Assistant's response} Use the images and scene graph as context and answer the following question . {Question Prompt}

Figure 45: CCoT Prompt.