

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

Benchmark	Task Objective	Visual Perspective	Question Perspective	Answer Type	Evaluator	#Questions (test)
MSVD-QA [40]	General Understanding	Exo	Exo	Predefined-Label	Accuracy	13K
MSRVTT-QA [40]	General Understanding	Exo	Exo	Predefined-Label	Accuracy	72K
Social-IQ [45]	Social Understanding	Exo	Exo	Multi-Choice	Accuracy	7.5K
Pano-AVQA [44]	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 [28]	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 [25]	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 Related Work

A.1 Ego-Exo Datasets and Tasks

Egocentric and exocentric views offer complementary information for understanding users and their environments. Early datasets like Charades-Ego [32] and LEMMA [14] introduced paired ego-exo data, while EgoExo4D [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 [36, 43]. 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, 48]. In parallel, cross-view knowledge transfer has been actively explored, with each perspective leveraged to improve the understanding of the other [47, 19, 41, 31]. Several studies have addressed viewpoint selection across perspectives by proposing methods for dynamically selecting informative views over time [26, 27]. Others have explored generating egocentric video from exocentric inputs using diffusion-based models [23, 42] or cropping third-person 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.

A.2 Visual Question Answering with LVLMS

Visual Question Answering (VQA) benchmarks test a model’s ability to interpret and reason over diverse visual content. Most existing VQA datasets are constructed from large-scale web-crawled data, typically consisting of images captured from fixed third-person cameras. MSVD-QA [40] and MSRVTT-QA [40] target general visual understanding through diverse question types, including what, how, when, where, and why. Pano-AVQA [44] evaluates spatial and audio-visual reasoning in panoramic 360° scenes, while Social IQ [45] focuses on social understanding by inferring the intentions and interactions of people within a scene. To support scenarios that require understanding from the user’s perspective, egocentric VQA datasets capturing first-person views have emerged. EgoVQA [8] evaluates first-person visual understanding by offering both egocentric and exocentric queries on first-person visual inputs. EgoSchema [28] 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 across diverse categories that reflect practical, real-world scenarios. Another line of work includes embodied QA benchmarks such as EmbodiedQA [5] and OpenEQA [25], where agents are required to navigate or interact with their environments to answer queries. Although numerous VQA datasets aim to evaluate LVLMS across diverse aspects, they cannot assess a model’s ability to seamlessly combine complementary visual information from paired ego and exo views (see Table 4).

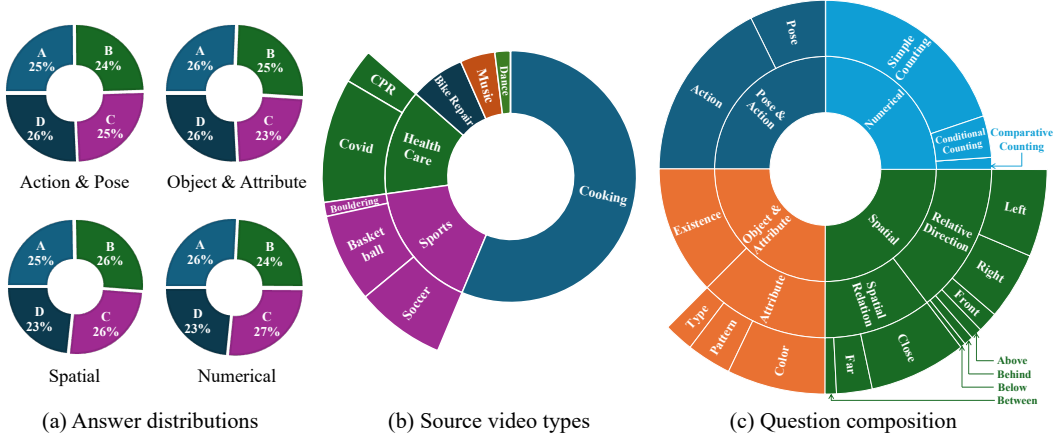


Figure 7: E3VQA statistics: (a) Distribution of correct answers across the four options (A–D), (b) Distribution of source video types used to construct E3VQA, and (c) Composition of question types within each category.

B E3VQA Benchmark Details

B.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.

B.2 Statistics

Figure 7 summarizes the statistics of the E3VQA benchmark. Figure 7(a) shows the distribution of correct answer choices across options (A–D) within each category. The uniform distribution of correct answers across the four options helps mitigate answer position bias. 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.

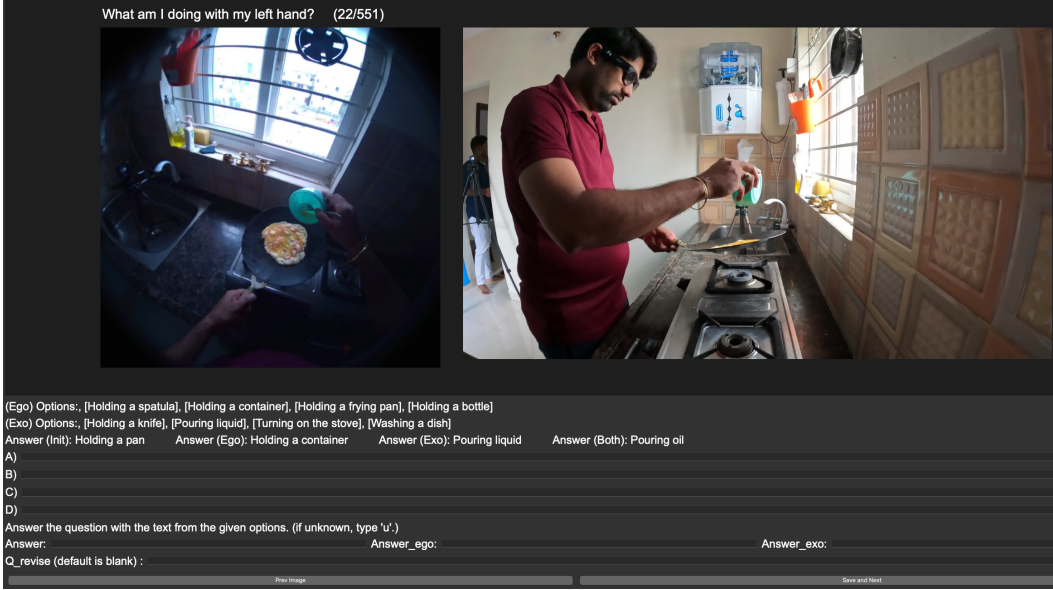


Figure 8: User interface provided for annotators during the human verification stage.

B.3 Details of Human Verification Stage

During the human verification stage, each of the four expert annotators is assigned to a specific question category and conducts verification using the user interface shown in Figure 8. Annotators initially utilized the view-specific responses generated in Section 2.3.2 to construct answer options. To supplement potentially redundant or low-quality responses, we generate two additional option sets, each containing four candidate options derived from the ego and exo images, respectively. Using these option sets, each annotator independently curates full question sets, including the question, correct answer, and distractor options, for their assigned category. After the initial construction, each annotator reviews subsets created by others to ensure consistency and clarity across the dataset, filtering out any ambiguous or low-quality instances.

C Experimental Details

C.1 LVLMs Overview 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.

C.2 CoT Baselines Overview

Building on its success in large language models (LLMs), CoT prompting has been extended to LVLMs to enhance inference-time reasoning. DDCoT [49] 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. CoCoT [46], 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. CCoT [29] helps understand the overall context of an image through scene graphs, where a scene graph is first generated via the LVLM and then incorporated into the prompt to enable compositional reasoning over objects, relations, and attributes. Despite their successes,

Table 5: Comparison of open-Source LVLMS: architecture (vision encoder and LLM) and egocentric data usage 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-Idetics2-8B	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 various 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 ± 1.50	54.93 ± 1.42	68.13 ± 0.81	73.73 ± 0.99	35.60 ± 1.11	53.00 ± 0.20	45.67 ± 0.58	48.33 ± 0.99	53.02
DDCoT [49]	47.87 ± 0.83	58.33 ± 2.64	68.47 ± 0.50	72.67 ± 1.42	35.33 ± 2.53	46.80 ± 2.12	50.67 ± 1.10	45.93 ± 0.95	53.26
CoCoT [46]	49.53 ± 0.81	57.27 ± 0.50	68.27 ± 1.14	72.53 ± 1.14	34.87 ± 1.55	47.93 ± 0.64	49.20 ± 1.91	46.27 ± 0.95	53.23
CCoT [29]	44.60 ± 1.91	58.40 ± 2.09	65.27 ± 0.64	73.80 ± 0.53	37.80 ± 3.30	50.00 ± 0.72	46.27 ± 1.33	48.80 ± 1.78	53.12
M3CoT (Ours)	45.87 ± 1.21	60.00 ± 0.35	70.60 ± 0.40	75.73 ± 0.76	35.07 ± 0.50	50.87 ± 0.70	50.80 ± 0.92	49.40 ± 0.72	54.79
InternVL3 - 8B									
Default	43.70 ± 3.25	54.90 ± 0.42	64.80 ± 0.85	70.30 ± 0.71	35.90 ± 2.12	45.20 ± 1.13	42.10 ± 2.97	46.60 ± 3.11	50.44
DDCoT [49]	48.10 ± 0.99	59.20 ± 4.24	67.20 ± 0.28	68.80 ± 1.13	34.60 ± 0.85	47.00 ± 0.00	46.30 ± 2.69	45.80 ± 1.41	52.13
CoCoT [46]	43.90 ± 0.14	58.20 ± 1.13	65.10 ± 0.42	68.40 ± 1.13	37.10 ± 1.84	48.70 ± 0.42	43.50 ± 2.97	43.40 ± 0.57	51.04
CCoT [29]	44.00 ± 0.85	55.00 ± 1.41	63.60 ± 0.57	68.30 ± 1.27	35.50 ± 0.42	51.20 ± 1.41	43.40 ± 0.57	44.70 ± 3.54	50.71
M3CoT (Ours)	45.50 ± 0.14	57.20 ± 0.00	68.20 ± 0.00	71.20 ± 0.00	37.60 ± 0.00	47.50 ± 0.71	53.80 ± 0.00	49.20 ± 0.00	53.78

their applicability to ego-exo multi-image contexts remains unexplored, raising an open challenge for extending CoT method to multi-image settings.

D Additional Experiments and Analysis of M3CoT

D.1 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 [50], the top-performing open-source model, and further evaluate the performance on InternVL3-8B. As shown in Table 6, 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 [38, 16, 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.

D.2 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

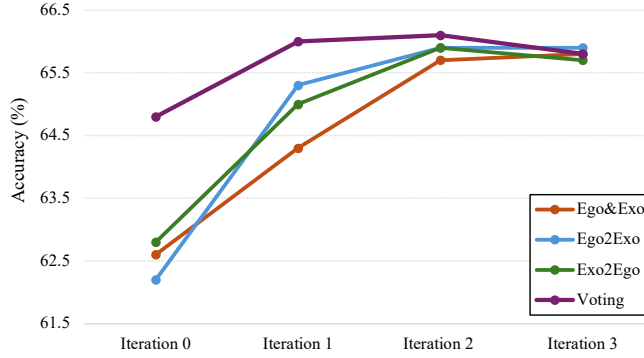


Figure 9: Performance across different perspectives and majority voting results over iteration steps.

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.

D.3 Qualitative Examples of 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 10, 11, and 12, 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.4 Additional Qualitative Examples of Different CoT Reasoning Processes

We provide additional examples that illustrate how our method improves reasoning compared to other CoT approaches (see Figure 13).

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 14–23. For view-specific response generation, we apply the prompts in Figure 24–31. For response-based filtering, we use the prompts shown in Figure 32 and 33. Finally, to generate four candidate options from either the ego or exo image, we use the prompts illustrated in Figure 34 and 35.

E.2 Prompt Templates for Experiments

The default system prompt and question prompt for E3VQA are shown in Figure 36. In addition, we present the prompts employed in the M3CoT. Scene graph generation prompts for each perspective are shown in Figure 37–39, and the prompts for scene graph refinement across agents are presented in Figure 40. For reference, the prompts used in other CoT methods are shown in Figure 41–43.

424 **F Limitations**

425 Despite its contributions, this work has several limitations. First, the E3VQA benchmark is solely
426 based on the EgoExo4D dataset, which may exhibit dataset bias and limited generalizability in
427 real-world visual assistant scenarios. Second, although the queries and answer options in E3VQA
428 are carefully crafted, they may not fully capture the diversity of natural language expressions and
429 user intents encountered in real-world interactions with visual AI assistants. Third, while recent
430 AI APIs offer a solution for scaling the benchmark, their use entails substantial financial costs.
431 Fourth, M3CoT introduces increased computational overhead due to its multi-step reasoning across
432 multiple perspectives, which may limit its applicability in resource-constrained scenarios. Finally,
433 since E3VQA is constructed from images rather than videos, the benchmark may not fully assess an
434 LVLM’s ability to capture temporal cues and motion dynamics, an aspect we leave for future work.

435 **G Ethics Statement**

436 This work has the potential to positively impact society by enhancing the capabilities of visual
437 assistants and embodied AI systems, particularly in scenarios that require comprehensive scene un-
438 derstanding from both egocentric and exocentric views. Such advancements may enhance human–AI
439 interaction and improve support in assistive technologies. However, the use of egocentric visual
440 data may raise important privacy concerns, especially in sensitive environments. We acknowledge
441 these risks and emphasize the importance of implementing safeguards and transparency mechanisms
442 in future deployments. As part of our commitment to responsible data use, we have obtained the
443 appropriate licenses from the contributing institutions for the use of the EgoExo4D dataset in this
444 research.





		<div>Ego2Exo Final Answer : D)</div>	<div>Scene graph: [{"objects": [{"id": "person1", "description": "person", "count": 1}]} ..., "summary": {"total_people": 1}}</div>
Ego View	Exo View	<div>Exo2Ego Final Answer : D)</div>	<div>Scene graph: [{"objects": [{"id": "person_1", "name": "person", "count": 3}]}], ...]</div>
<div>Q: How many people are in the scene? A) 2 C) 1</div>		<div>Ego&Exo Final Answer : B)</div>	<div>Scene graph: [{"objects": [{"id": "person1", "perspective": "image1"}, {"id": "person2", "perspective": "image1"}, {"id": "person3", "perspective": "image1"}, {"id": "person4", "perspective": "image2"}]}]</div>
		<div>Ego2Exo Final Answer : D)</div>	<div>Scene graph: [{"objects": [{"id": "person1", "name": "person", "attributes": {"standing": true}}], "relationships": [{"subject": "person1", "relation": "in", "object": "scene"}]}</div>
Ego View	Exo View	<div>Exo2Ego Final Answer : D)</div>	<div>Scene graph: [{"objects": [{"id": "person1", "attributes": ["wearing glasses", "wearing a white shirt"]}, {"id": "person2", "attributes": "standing"}, {"id": "person3", "attributes": "sitting"}], ..., "summary": "There are three people in the scene."}</div>
<div>Q: How many people are in the scene? A) 2 C) 1</div>		<div>Ego&Exo Final Answer : B)</div>	<div>Scene graph: "objects": [{"id": "person1", "perspective": "image1"}, {"id": "person2", "perspective": "image1"}, {"id": "person3", "perspective": "image1"}, {"id": "user", "perspective": "image2"}], "total_people": 4}</div>

Figure 10: 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.


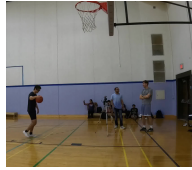


		Ego2Exo Final Answer : A)	Scene graph: [{"objects": [{"id": "person_light_blue_shirt", "attributes": {"shirt_color": "light blue", "arm_position": "hands resting by the sides"}}, ... , {"id": "person_crossing_arms", ...}]}
Ego View	Exo View	Exo2Ego Final Answer : C)	Scene graph: [{"objects": [{"name": "person", "attributes": "wearing light blue shirt"} , ... , {"name": "arms", "relation": "crossed", "target": "person"}]}
Q: How is the person in the light blue shirt positioned? A) Hands resting by the sides B) Sitting on a bench C) Crossing arms D) Leaning on a ladder		Ego&Exo Final Answer : C)	Scene graph: [{"objects": [{"id": "person1", "description": "person in light blue shirt", "attributes": {"shirt_color": "light blue", "position": "standing", "pose": "crossing arms"}, ... }]}]
		Ego2Exo Final Answer : B)	Scene graph: [{"objects": [{"id": "person", "description": "person playing violin"}, ... , {"id": "tripod", "relation": "to the right of", "target": "person"}, ...]}]
Ego View	Exo View	Exo2Ego Final Answer : C)	Scene graph: [{"objects": [{"id": "person", "relation": "playing", "target": "violin"}, ... , {"id": "tripod", "relation": "in front of", "target": "person"}]}]
Q: What object is on the right of the person playing violin? A) Music stand B) Tripod C) Light switch D) Trash bin		Ego&Exo Final Answer : C)	Scene graph: [{"objects": [{"id": "tripod", "relation": "to the right of", "target": "person"}, ... , {"id": "light_switch", "relation": "to the right of", "target": "person"}]}]

Figure 11: 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.

		<p>Ego2Exo Final Answer : A)</p> <p>Scene graph: [{"objects": [{"name": "man", "attributes": ["in light green shirt", "sitting"], "relation": "holding", "target": "swab", "hand": "left"}, {"name": "timer", "relation": "on", "target": "table"}]}]</p>
<p>Q: What is the man in a light green top holding in his left hand?</p> <p>A) Box B) Instruction manual C) Timer D) Phone</p>		<p>Exo2Ego Final Answer : C)</p> <p>Scene graph: [{"objects": [{"name": "man", "attributes": ["sitting", "light green shirt"], "relation": "holding", "target": "timer"}], ...}]</p>
		<p>Ego2Exo Final Answer : C)</p> <p>Scene graph: [{"objects": [{"name": "frying pan", "relation": "far", "target": "window"}, {"name": "gas cylinder", "relation": "near", "target": "window"}], ...}]</p>
<p>Q: What is the object closest to the window?</p> <p>A) Frying pan B) Cutting board C) Gas cylinder D) Knife</p>		<p>Exo2Ego Final Answer : A)</p> <p>Scene graph: [{"objects": [{"name": "window", ...}, {"name": "frying pan", "relation": "closest to", "target": "window"}]}]</p>
		<p>Ego&Exo Final Answer : C)</p> <p>Scene graph: [... , {"relationships": [{"source": "frying_pan", "target": "window", "type": "far_from"}, {"source": "gas_cylinder", "target": "window", "type": "near"}]}]</p>

Figure 12: Qualitative examples of answers and reasoning processes generated by different perspectives. The scene graph from the Exo2Ego perspective demonstrates a strong capability to capture the information necessary for answering questions grounded in the ego view alone.

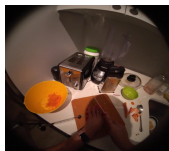

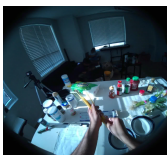

		<p>DDCoT</p> <p>Final Answer : A)</p> <p>Sub question 1: What is to the right of the sink in the second image? Sub Answer 1: In the second image, the stove is to the right of the sink Sub question 2: Which of the answer choices is to the right of the sink? Sub Answer 2: The cutting board is on the counter to the right of the sink</p>
Ego View	Exo View	<p>CCoT</p> <p>Final Answer : A)</p> <p>Scene graph: [{"object": "person", "relation": "to the right of", "target": "sink"}, ... { "object": "cutting board", "relation": "to the right of", "target": "sink" }]</p>
<p>Q: What object is to the right of the sink, relative to the person?</p> <p>A) Cutting board B) Stove C) Blender D) Toaster</p>		<p>M3CoT</p> <p>Final Answer : B)</p> <p>Scene graph: [{"objects": [{"id": "person", "relation": "near", "target": "sink"}, ... , {" id": "stove", ... , "relation": "to the right of", "target": "sink" }, ...]]</p>
		<p>DDCoT</p> <p>Final Answer : A)</p> <p>Sub question 1 : Where is the blue plate located? Sub Answer 1 : The blue plate is located on the kitchen island. Sub question 2 : What cooking utensil is on the blue plate? Sub Answer 2 : The cooking utensil on the blue plate is a fork.</p>
Ego View	Exo View	<p>CCoT</p> <p>Final Answer : A)</p> <p>Scene graph: [{"name": "plate", "attribute": "blue", "relation": "on", "target": "island"}, { "name": "fork", "relation": "on", "target": "plate" }]</p>
<p>Q: Which cooking utensil is on the blue plate positioned on the kitchen island?</p> <p>A) Fork B) Pillar C) Spatula D) Knife</p>		<p>M3CoT</p> <p>Final Answer : C)</p> <p>Scene graph: [{"objects": [{"name": "island", "attribute": "kitchen"}, {"name": "plate", "attribute": "blue", "relation": "on", "target": "island" }, {"name": "spatula", "relation": "on", "target": "plate" }, ...]]</p>

Figure 13: Qualitative examples of answers and reasoning processes generated by different prompting methods.

Egocentric Single-View QA Generation Prompt
<p>{Ego Image}</p> <p>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.</p> <p>{Category-wise Prompt}</p> <p>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.</p>

Figure 14: Egocentric single-view QA generation prompt.

Egocentric Single-View QA Generation Prompt: Action & Pose
<p>Instructions :</p> <p>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’.</p> <p>Question Categories & Templates:</p> <p>Actions (What am I doing?)</p> <ul style="list-style-type: none"> – What am I doing? – What am I doing with my [body part]? <p>Body Posture (How am I positioned?)</p> <ul style="list-style-type: none"> – How is my body positioned? – How am I sitting /standing /lying? – What is my posture? <p>Gestures (What movement am I making?)</p> <ul style="list-style-type: none"> – What am I doing with my hands? – What gesture am I making? – How am I moving my arms/legs/head? <p>Examples:</p> <p>Q: How is my body positioned? A: Sitting cross-legged</p> <p>Q: What am I doing with my left hand? A: Holding a book</p> <p>Q: What gesture am I making? A: Waving</p>

Figure 15: Egocentric single-view QA generation prompt: Action & Pose.

Egocentric Single-View QA Generation Prompt: Object & Attribute
<p>Instructions :</p> <p>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.</p> <p>The answer must be a noun or noun phrase , avoiding overly generic responses such as something or object.</p> <p>Question Categories & Templates:</p> <p>Object Identification (What am I interacting with?)</p> <ul style="list-style-type: none"> - What am I holding? - What object is on the table beside me? - Which item am I picking up? <p>Object Attributes (What does it look like?)</p> <ul style="list-style-type: none"> - What color is the shirt I am wearing? - What pattern is on my jacket? - What type of shoes am I wearing? <p>Examples:</p> <p>Q: What color is the shirt I am wearing?</p> <p>A: Navy blue</p> <p>Q: Which object am I holding in my right hand?</p> <p>A: A small notebook</p> <p>Q: What pattern does my sweater have?</p> <p>A: Checkered pattern</p>

Figure 16: Egocentric single-view QA generation prompt: Object & Attribute.

Egocentric Single-View QA Generation Prompt: Spatial
<p>Instructions:</p> <p>Each question must focus on the spatial relationships between me and objects in my surroundings.</p> <p>The answer must be a specific object or location descriptor (e.g., coffee cup, bookshelf, under the table).</p> <p>Do not generate QA pairs with overly generic answers.</p> <p>Question Categories & Templates:</p> <p>Object Proximity (What is closest or farthest?):</p> <ul style="list-style-type: none"> - What object is closest to me? - Which object is the farthest from me? - What is the nearest object to my [body part]? <p>Relative Positioning (Where are objects located?):</p> <ul style="list-style-type: none"> - 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]? <p>Examples:</p> <p>Q: What object is closest to my left hand?</p> <p>A: Coffee cup</p> <p>Q: Which object is the farthest from me?</p> <p>A: Bookshelf</p> <p>Q: What object is on my right side?</p> <p>A: Tissue</p>

Figure 17: Egocentric single-view QA generation prompt: Spatial.

Egocentric Single-View QA Generation Prompt: Numerical
<p>Instructions:</p> <p>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.</p> <p>Question Categories & Templates:</p> <p>Counting People (How many people are around me?)</p> <ul style="list-style-type: none"> – How many people are in the image excluding me? – How many individuals are facing the same direction as I am? <p>Counting Objects (How many things are near or with me?)</p> <ul style="list-style-type: none"> – How many [objects] am I holding? – How many [items] are on the table beside me? <p>Quantitative Comparisons (How do the numbers compare to what I have?)</p> <ul style="list-style-type: none"> – 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? <p>Examples:</p> <p>Q: How many people are in the image excluding me? A: 3</p> <p>Q: How many more bowls are on my table compared to the table behind me? A: 2</p> <p>Q: How many apples am I holding? A: 3</p>

Figure 18: 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 19: 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 20: Exocentric single-view QA generation prompt: Action & Pose.

Exocentric Single-View QA Generation Prompt: Object & Attribute
<p>Instructions:</p> <p>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').</p> <p>Questions should reference people or objects by descriptors (e.g., 'the woman in the white top', 'the man with the striped shirt').</p> <p>The answer must be a noun or noun phrase, avoiding overly generic responses such as 'something' or 'object'.</p> <p>Question Categories & Templates:</p> <p>Object Identification (What is present?)</p> <ul style="list-style-type: none"> - What is the man with the striped shirt holding? - What object is placed on the table? - Which item is the woman wearing blue top picking up? <p>Object Attributes (What does it look like?)</p> <ul style="list-style-type: none"> - 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? <p>Examples:</p> <p>Q: What color is the top worn by the woman holding the towel? A: White</p> <p>Q: Which object is the man in the black shirt holding in his right hand? A: Smartphone</p> <p>Q: What pattern does the sweater worn by the person holding a cup have? A: Checkered pattern</p>

Figure 21: Exocentric single-view QA generation prompt: Object & Attribute.

Exocentric Single-View QA Generation Prompt: Spatial
<p>Instructions:</p> <p>Each question must explicitly reference an object's or a person's spatial relationship within the scene.</p> <p>The answer must be a specific object or location descriptor (e.g., scissors, frying pan, under the table).</p> <p>Do not generate QA pairs with overly generic answers.</p> <p>Question Categories & Templates:</p> <p>Object Proximity (What is closest or farthest?)</p> <ul style="list-style-type: none"> - 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]? <p>Relative Positioning (Where are objects located?)</p> <ul style="list-style-type: none"> - 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]? <p>Spatial Relations (How are objects arranged?)</p> <ul style="list-style-type: none"> - 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? <p>Examples:</p> <p>Q: What is the object on the far right of the desk? A: Scissors</p> <p>Q: Which cookware is closest to the woman wearing a striped shirt? A: Frying pan</p> <p>Q: What object is placed directly in front of the man wearing a cap? A: Backpack</p> <p>Q: What object is placed underneath the table? A: Storage box</p>

Figure 22: Exocentric single-view QA generation prompt: Spatial.

Exocentric Single-View QA Generation Prompt: Numerical
<p>Instructions :</p> <p>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.</p> <p>Question Categories & Templates:</p> <p>Counting People (How many are there?)</p> <ul style="list-style-type: none"> – How many people are in the scene? – How many individuals are facing the camera? <p>Counting Objects (How many things are visible?)</p> <ul style="list-style-type: none"> – How many objects is [person descriptor] holding? – How many items are on the table? <p>Quantitative Comparisons (How do the numbers compare?)</p> <ul style="list-style-type: none"> – 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? <p>Examples:</p> <p>Q: How many people are in the scene? A: 3</p> <p>Q: How many objects is the woman in the striped shirt holding? A: 2</p> <p>Q: How many oranges are placed on the table? A: 5</p>

Figure 23: Exocentric single-view QA generation prompt: Numerical.

View-Specific Response Expansion Prompt: Ego View
<p>{Ego Image}</p> <p>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.</p> <p>{Category-wise Prompt}</p> <p>Requirements: Each answer option should be a single word or a short phrase. Follow the provided format strictly.</p> <p>Q: {Question}</p>

Figure 24: View-specific response expansion prompt: Ego view.

View-Specific Response Expansion Prompt: Exo View
<p>{Exo Image}</p> <p>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.</p> <p>{Category-wise Prompt}</p> <p>Requirements: Each answer should be a single word or a short phrase. Follow the provided format strictly.</p> <p>Q: {Question}</p>

Figure 25: View-specific response expansion prompt: Exo view.

View-Specific Response Expansion Prompt: Both Views
<p>{Ego Image}</p> <p>{Exo Image}</p> <p>You are provided with two visual inputs in sequence, each captured from a different perspective:</p> <ol style="list-style-type: none"> 1. The view from the camera worn by the user ('I'). 2. The view captured by an external camera observing the user ('I'). <p>These two images capture the same event at the same time. Based on the visual inputs, generate the best possible answer.</p> <p>{Category-wise Prompt}</p> <p>Requirements:</p> <p>Each answer should be a single word or a short phrase. Follow the provided format strictly.</p> <p>Q: {Question}</p>

Figure 26: View-specific response expansion prompt: Both views.

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

Figure 27: View-specific response expansion prompt: text only.

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

Figure 28: View-specific response expansion prompt: Action & Pose.

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

Figure 29: View-specific response expansion prompt: Object & Attribute.

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

Figure 30: View-specific response expansion prompt: Spatial.

View-Specific Response Expansion Prompt: Numerical
<p>Instructions:</p> <p>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.</p> <p>Output format:</p> <p>Q: How many people are in the image excluding me?</p> <p>A: 3</p> <p>Q: How many people are in the scene?</p> <p>A: 3</p>

Figure 31: View-specific response expansion prompt: Numerical.

Response-Based Question Filtering Prompt 1
<p>Here is the question: '{Question}'.</p> <p>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.</p>

Figure 32: Response-based question filtering prompt (1).

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

Figure 33: Response-based question filtering prompt (2).

Option Generation Prompt: Ego
<p>{Ego Image}</p> <p>You are given a visual input from a camera worn by the user (referred to as 'I').</p> <p>Based on the following question and answer, generate four multiple-choice options.</p> <p>Question: {Question}</p> <p>Answer: {answer_ego}</p> <p>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:</p> <p>Options:</p> <p>[Option1]</p> <p>[Option2]</p> <p>[Option3]</p> <p>[Option4]</p>

Figure 34: Option generation prompt: Ego.

Option Generation Prompt: Exo

`{Exo Image}`

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: `{Question}`

Answer: `{answer_exo}`

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 35: Option generation prompt: Exo.

System Prompt & Question (Instruction) Prompt
<p>System Prompt</p> <p>You are a helpful assistant. You are provided with two visual inputs in sequence, each captured from a different perspective:</p> <ol style="list-style-type: none"> 1. The view from the camera worn by the user ('I'). 2. The view captured by an external camera observing the user ('I'). <p>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.</p> <p>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.</p> <p>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.</p> <p>Question (Instruction) Prompt</p> <p>{Ego Image} {Exo Image} {Question}</p> <p>Only one option is correct. Present the answer in the form X).</p>

Figure 36: System Prompt and Question(Instruction) Prompt.

M3CoT Prompts - Ego2Exo Perspective
<p>Scene graph generation phase (Ego2Exo)</p> <p>Task: For the provided image and its associated question, generate a scene graph in JSON format that includes the following:</p> <ol style="list-style-type: none"> 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. <p>Just generate the scene graph in JSON format. Do not say extra words.</p> <p>{Ego Image} {Question Prompt}</p> <hr/> <p>Scene graph refinement phase (Ego2Exo)</p> <p>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:</p> <ol style="list-style-type: none"> 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. <p>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.</p> <p>{Exo Image} {Question Prompt} {Assistant's response(Ego-only SG)}</p> <hr/> <p>Initial question response phase (Ego2Exo)</p> <p>Use the images and the refined scene graph as context and answer the following question.</p> <p>{Ego Image} {Exo Image} {Question Prompt} {Assistant's response(Refined SG)}</p>

Figure 37: M3CoT prompt (1).

M3CoT Prompts - Exo2Ego Perspective
<p>Scene graph generation phase (Exo2Ego)</p> <p>Task: For the provided image and its associated question, generate a scene graph in JSON format that includes the following:</p> <ol style="list-style-type: none"> 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. <p>Just generate the scene graph in JSON format. Do not say extra words.</p> <p>{Ego Image} {Question Prompt}</p> <hr/> <p>Scene graph refinement phase (Exo2Ego)</p> <p>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:</p> <ol style="list-style-type: none"> 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. <p>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.</p> <p>{Exo Image} {Question Prompt} {Assistant's response(Exo-only SG)}</p> <hr/> <p>Initial question response phase (Exo2Ego)</p> <p>Use the images and the refined scene graph as context and answer the following question.</p> <p>{Ego Image} {Exo Image} {Question Prompt} {Assistant's response(Refined SG)}</p>

Figure 38: M3CoT prompt (2).

M3CoT Prompts - Ego&Exo Perspective
<p>Scene graph generation phase (Ego&Exo)</p> <p>Task :</p> <p>Using the provided two images and their associated question , generate a unified scene graph in JSON format that includes the following :</p> <ol style="list-style-type: none"> 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 . 4. Ensure that objects and relationships from both perspectives are appropriately aligned , integrated and refined to provide a complete scene representation . <p>Just generate the unified scene graph in JSON format. Do not say extra words.</p> <p>{Ego Image} {Exo Image} {Question Prompt}</p> <hr/> <p>Initial question response phase (Ego&Exo)</p> <p>Use the images and the unified scene graph as context and answer the following question .</p> <p>{Ego Image} {Exo Image} {Question Prompt} {Assistant's Response(Ego&Exo SG)}</p>

Figure 39: M3CoT prompt (3).

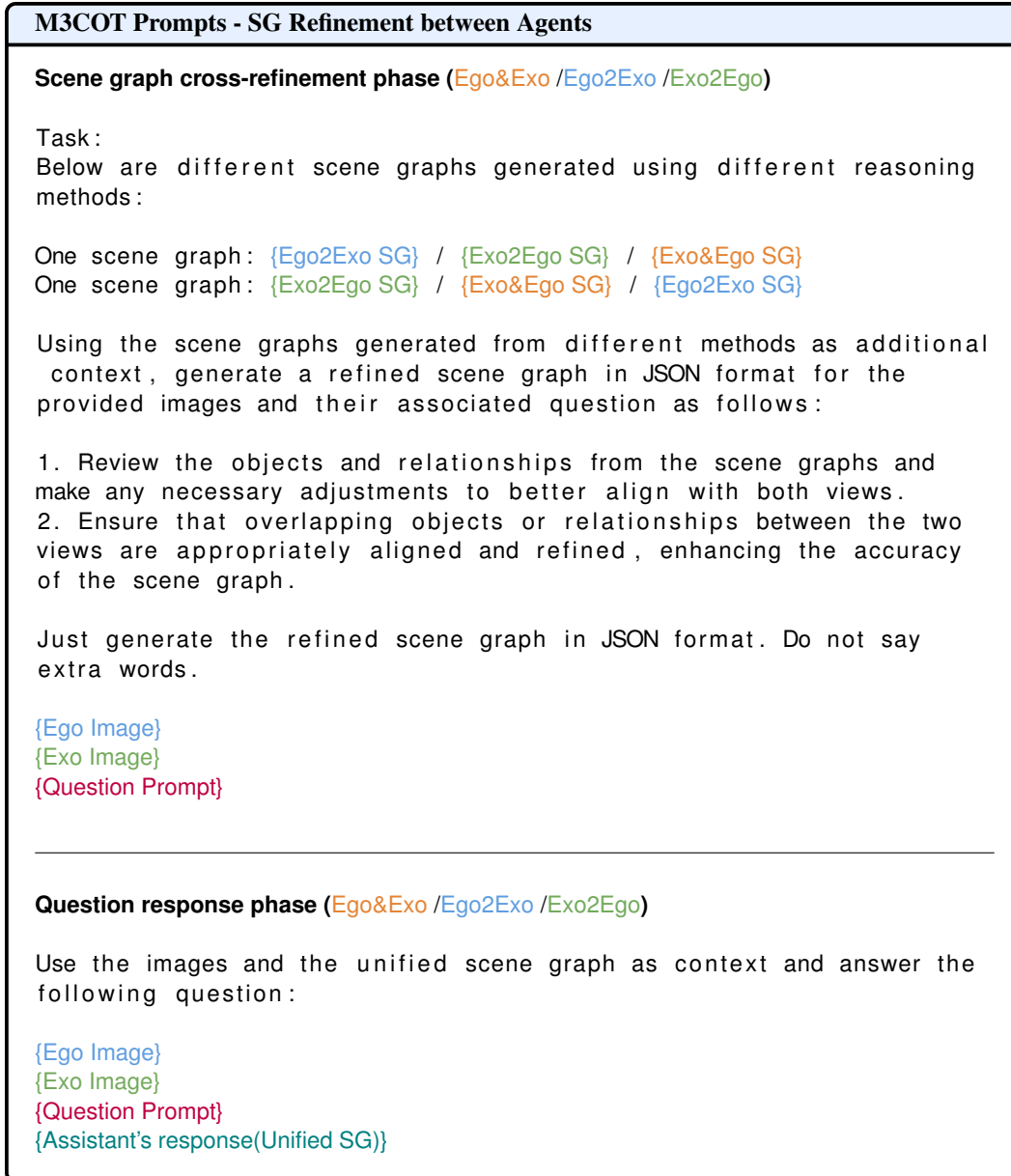


Figure 40: M3CoT prompt (4).

Other CoT Prompts - DDCoT

For the provided images and their associated question, think step-by-step about the preliminary knowledge required to answer the question. Deconstruct the problem as completely as possible into necessary sub-questions.

Then, with the aim of helping humans answer the original question, attempt to answer those sub-questions.

The expected answering format is as follows:

Sub-questions:

- <sub-question 1>
- <sub-question 2>
- ...

Sub-answers:

- <sub-answer 1>
- <sub-answer 2>
- ...

{Question Prompt}

Context: {Assistant's response}

Give your answer of the question according to the sub-questions and sub-answers.

{Question Prompt}

Figure 41: DDCoT Prompt.

Other CoT Prompts - CoCoT

Please tell me the similarities and differences of these two images, and answer to the question.

{Question Prompt}

Figure 42: CoCoT Prompt.

Other CoT Prompts - CCoT
<p>For the provided images and their associated question , generate a scene graph in JSON format that includes the following :</p> <ol style="list-style-type: none">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 . <p>Just generate the scene graph in JSON format . Do not say extra words .</p> <p>{Question Prompt}</p> <hr/> <p>Scene Graph: {Assistant's response}</p> <p>Use the images and scene graph as context and answer the following question .</p> <p>{Question Prompt}</p>

Figure 43: CCoT Prompt.

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, Sibao 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 Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024.
- [16] 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.
- [17] 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.
- [18] 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.

- [19] 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.
- [20] 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.
- [21] 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.
- [22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] Gemini Team. Gemini: A family of highly capable multimodal models, 2024.
- [35] 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.
- [36] 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.

- [37] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024.
- [38] 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.
- [39] Yiqi Wu, Xiaodan Hu, Ziming Fu, Siling Zhou, and Jiangong Li. Gpt-4o: Visual perception performance of multimodal large language models in piglet activity understanding. *arXiv preprint arXiv:2406.09781*, 2024.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] 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 Wenhao Wang. Internvl3: 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.