

From Observation to Understanding: Front-Door Adjustments with Uncertainty Calibration for Enhancing Egocentric Reasoning in LVLMs

Shenshen Li¹, Wenxin Meng¹, Lei Wang³, Hao Yang⁴, Chong Peng⁴,
Peng Yan⁴, Fumin Shen¹, Jingkuan Song^{1,2}, Heng Tao Shen², Xing Xu^{1,2*}

¹University of Electronic Science and Technology of China

²School of Computer Science and Technology, Tongji University

³Salesforce AI Research, ⁴Meituan

Abstract

Recent progress in large vision-language models (LVLMs) has shown substantial potential across a broad spectrum of third-person tasks. However, adapting these LVLMs to egocentric scenarios remains challenging due to their third-person training bias. Existing methods that adapt LVLMs for first-person tasks often overlook critical agent-environment interactions, limiting their ability to perform egocentric reasoning. To address these challenges, we propose a novel zero-shot paradigm termed *Front-Door Adjustments with Uncertainty Calibration (FRUIT)* to enhance the egocentric reasoning abilities of LVLMs by simulating human causal reasoning. Specifically, the FRUIT operates in two stages: *observation and understanding*. Unlike conventional prompting techniques, we formalize egocentric reasoning using a structural causal model. Then, we ground interaction regions and expand them into hierarchical visual cues, augmented with corresponding captions, to form the initial observations. To reduce noise in these observations, we employ uncertainty calibration to filter out unreliable information. These refined observations as mediators are then incorporated into the prompt template, guiding the model to understand semantics from a first-person perspective. Extensive experiments conducted on the EgoThink benchmark demonstrate that our FRUIT method consistently enhances the performance of existing LVLMs on six distinct tasks. Our code is available at <https://github.com/Mrshenshen/FRUIT>.

1 Introduction

Recently, large vision-language models (LVLMs) (Bai et al., 2023; Zhu et al., 2024) have made significant progress in third-person scene understanding tasks (Li et al., 2023b, 2024a). Meanwhile, the growing research interest in egocentric tasks has driven the expansion of LVLMs into domains such

* Corresponding authors.

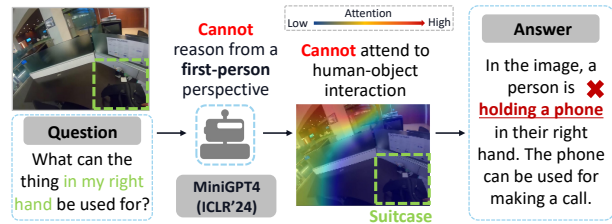


Figure 1: Illustrative examples of existing problems: the recent LVLm MiniGPT4 (Zhu et al., 2024) fails to accurately focus on human-object interactions and recognize the object held in the right hand as a phone rather than a suitcase. Such results exemplify the limitations of existing LVLMs in egocentric reasoning.

as intelligent agents (Liu et al., 2024b) and wearable devices (Zhang et al., 2024b).

However, the shift towards egocentric tasks, which requires models to interpret the dynamic interactions between agents and their environment from the subjective viewpoint of the agent, exposes limitations inherent in current LVLMs. These limitations stem from a critical mismatch: LVLMs typically trained on third-person data, lack the mechanisms necessary to prioritize agent-environment interactions that are crucial for egocentric reasoning. Such reasoning is essential for applications like intelligent agents and wearable devices.

Existing approaches generally adopt two primary strategies: 1) fine-tuning models on egocentric datasets (Kukleva et al., 2024), typically limited to specific egocentric tasks; 2) applying prompt engineering techniques (Lin et al., 2024). However, both paradigms share a fundamental limitation: they fail to adequately explore the underlying causal relationships from a first-person perspective, *i.e.*, how actions, events, and contextual factors interact to produce specific outcomes. This drawback leads to suboptimal egocentric reasoning ability (Cheng et al., 2024) across a broad spectrum of first-person tasks. (Wang et al., 2023; Li et al., 2022). As shown in Figure 1, the recent LVLm MiniGPT4 (Zhu et al., 2024) fails to accurately rec-

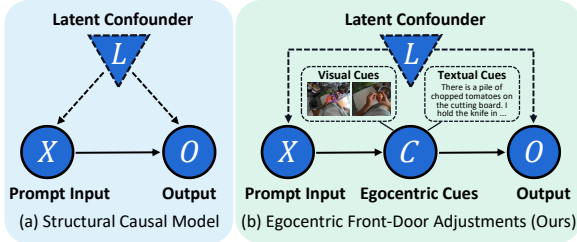


Figure 2: Pipeline representation through Causal Modeling: (a) The Structural Causal Model including prompt input X , Output O and latent confounder L ; (b) The proposed Egocentric Front-Door Adjustment that helps model better reason from a first-person perspective.

ognize object interactions and understand spatial relationships from a first-person perspective, as it does not prioritize active objects in the agent’s view, missing key contextual cues for effective reasoning.

To address this problem, inspired by the two-stage reasoning of human cognition (Clark, 2013), *i.e.*, first gathering interaction-centric observations and then deriving contextual understanding, we propose a novel paradigm termed *Front-Door Adjustments with Uncertainty Calibration (FRUIT)* for enhancing the egocentric reasoning abilities of LVLMs. The FRUIT consists of two key stages: 1) *Mediated Observation*: Introducing hierarchical egocentric interaction cues as causal mediators, such as human-object affordance visual grounding, effectively circumventing the confounding biases from third-person training. 2) *Contextual Understanding*: The model refines its understanding by focusing on key observations, making more accurate and context-driven reasoning decisions.

Specifically, as illustrated in Figure 2(a), instead of conventional prompting methods (Lin et al., 2024), we first capture the causal relationships between input prompts and outputs through a Structural Causal Model (SCM). Building upon this, as shown in Figure 2(b), we propose an Egocentric Front-Door Adjustment to introduce hierarchical multi-modal egocentric cues as *mediated observations* within the LVLm reasoning pipeline, using the grounding model (Liu et al., 2023c). Moreover, to reduce the influence of noise in cues, we develop an Uncertainty Calibration Mechanism to select reliable egocentric cues. Following the front-door criterion (Shanmugam, 2001), such cues are treated as mediator variables C between prompt X and output O , with latent confounder edges L omitted for simplicity. This framework enables the model to focus its *reasoning* on interaction-centric observations while maintaining a comprehensive *under-*

standing of the global semantic context. We evaluate our FRUIT method on the EgoThink benchmark dataset (Cheng et al., 2024). Extensive experimental results prove that our method can consistently improve the egocentric reasoning ability of existing LVLMs on six various egocentric tasks.

Our contributions can be summarized as follows: 1) We propose a novel paradigm named *Front-Door Adjustments with Uncertainty Calibration (FRUIT)* to improve the egocentric reasoning ability of LVLMs by simulating the two-stage reasoning process in human cognition: observation and understanding. 2) We propose an Egocentric Front-Door Adjustment (EFDA) scheme to introduce mediated observation, which aims to guide the model in focusing on the human-object interaction, thus reasoning from a first-person perspective. 3) We design an Uncertainty Calibration Mechanism (UCM) to effectively filter out unreliable cues by modeling their inherent uncertainty.

2 Related Work

Large Vision-Language Models. Initially confined to natural language processing, LLMs have recently extended their capabilities to multi-modal tasks, particularly through the development of Large Vision-Language Models (LVLMs) (Chen et al., 2024; Su et al., 2023). These models typically undergo a two-stage training process: pre-training for feature alignment and instruction-based fine-tuning, enabling them to achieve strong performance across tasks like visual question answering (Wang et al., 2024), object detection (Li et al., 2024b), and image segmentation (Li et al., 2024c). While most LVLm evaluations are designed based on third-person data, emerging benchmarks prioritize egocentric assessments (Cheng et al., 2024). In this work, we aim to enhance egocentric processing and reasoning abilities of LVLMs.

Causal Mechanism. Causal mechanisms (Rohekar et al., 2023; Zhang et al., 2024a) in LVLMs have recently garnered significant attention due to their potential to address complex reasoning tasks. To refine the reasoning process within LVLMs, several methods have utilized existing causal frameworks, such as Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) and Structural Causal Model (SCM) (Schölkopf et al., 2012), which aim to explore the causal relationships. For example, SCM has provided a structured foundation to dissect causal relationships (Rohekar et al., 2023), allow-

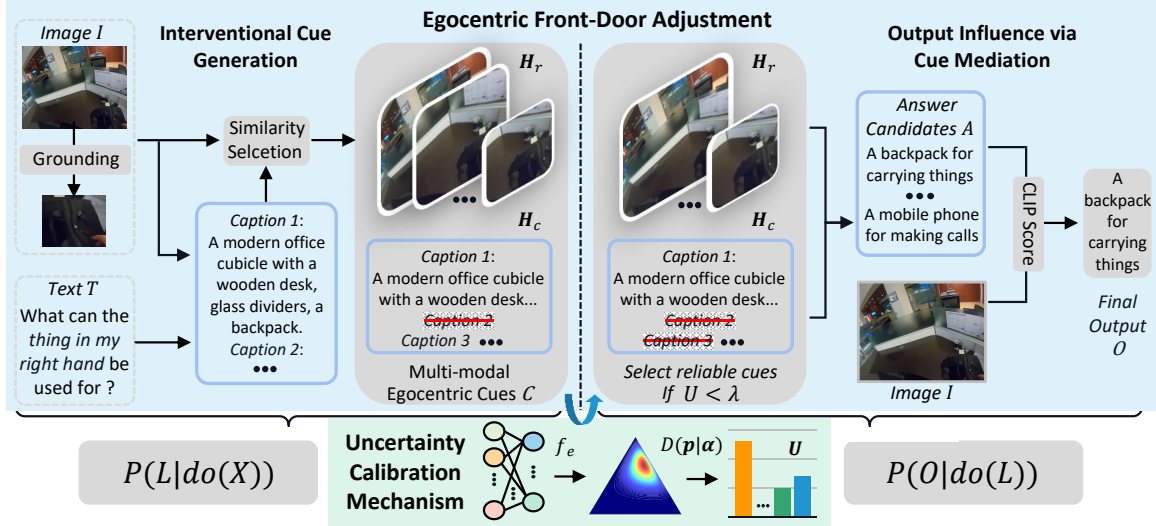


Figure 3: The overall framework of the proposed FRUIT method. It includes two components: Egocentric Front-Door Adjustment (EFDA) that contains Interventional Cue Generation (ICG) and Output Influence via Cue Mediation (OICM), and Uncertainty Calibration Mechanism (UCM). The “Caption” denotes the deletion of caption.

ing LVLMs to isolate causal factors that contribute directly to the reasoning process. In this paper, we model the reasoning pipeline of LVLMs in first-person scenarios using an SCM.

Uncertainty-based Learning. Uncertainty-based learning addresses the challenge of quantifying prediction confidence, with uncertainty typically classified into epistemic and aleatoric types (Kendall and Gal, 2017). Aleatoric uncertainty arises from inherent data noise, while epistemic uncertainty reflects a lack of knowledge about the model structure. Methods like Bayesian networks (Gal and Ghahramani, 2016) and Subjective Logic (Jsang, 2018) are used to quantify epistemic uncertainty, often employing Dempster-Shafer theory (Yager and Liu, 2008). To effectively filter out the unreliable egocentric cues, we design an Uncertainty Calibration Mechanism to model their uncertainty.

3 Proposed Method

Preliminary Overview. Our primary goal is to enhance the egocentric reasoning of LVLMs to generate precise responses in first-person contexts. Given an input prompt X including an image I and text T , we first employ a grounding model to derive hierarchical visual egocentric cues H_r and textual cues H_c , which can be considered as mediated observations. After filtering noise in the observation, we incorporate observations into the prompt template following the front-door criterion to obtain a set of answer candidates A . Finally, we compute the CLIP score between each candidate in A and the input image, selecting the candidate

with the highest score as the final output O .

3.1 Background of Structural Causal Model

Structural Causal Model. An SCM (Shanmugam, 2001) aims to capture causal mechanisms between variables. Following prior works (Rohekar et al., 2023), an SCM can be formalized as a Directed Acyclic Graph (DAG), denoted as $G = (V, E)$, where V represents the set of variables and E denotes the directed edges that signify direct causal relationships between these. In this context, the variables in DAG are topologically ordered, ensuring each variable precedes its causal descendants, based on fundamental assumptions outlined below:

Definition 1 (Causal Markov). A variable in a causal graph that satisfies the Causal Markov Condition is independent of all other variables, excluding its direct effects, given its direct causes.

Definition 2 (Faithfulness). A distribution is faithful to a graph if and only if all independence relations that hold within the distribution are also captured by the corresponding graphical structure.

Pipeline Modeling through SCM. We use an SCM to capture the causal relationships and dependencies between input prompts and outputs in LVLMs for egocentric tasks. Specifically, as illustrated in Figure 2(b), the SCM is represented as a directed acyclic graph $G = (V, E)$. We define the components of our SCM as follows: 1) X : The input prompt, comprising visual and textual elements from the egocentric environment. 2) O : The LVM output in response to the input prompt X , such as scene descriptions or object identifica-

tion. 3) L : The latent confounder, representing latent variables that simultaneously affect both input X and output O , potentially introducing biases or noise in the learned relationships. Moreover, in Figure 2(b), the notation $X \rightarrow O$ denotes a direct causal relationship, illustrating that the varying multi-modal input prompts X in egocentric context lead to predictable adjustments in outputs O . The arrows $L \rightarrow X$ and $L \rightarrow O$ denote confounding influences from unobserved variables.

3.2 Egocentric Front-Door Adjustment

Due to the inaccessibility of the confounding variable L , the back-door adjustment is not feasible. Therefore, we follow the front-door criterion to integrate hierarchical multimodal egocentric cues as mediated variables, simulating the human cognition observation stage. It facilitates LVLMs focusing on human-centric interactions rather than contextual intentions on all semantics.

Specifically, as illustrated in Figure 2(b), the proposed EFDA incorporates egocentric cues as a mediator C between the prompt X and output O , which serves as a bridge to account for the indirect influence of the prompt on the answer. Moreover, following the front-door criterion (Shanmugam, 2001), we ignore the confounder of L with other variables. To quantify the causal effect between X and O , we employ the causal intervention through the do-operation, which is formulated as follows:

$$P(O|do(X)) = P(O|do(C))P(C|do(X)), \quad (1)$$

where the causal effect of $P(O|do(X))$ between X and O can be split into: $P(C|do(X))$ represents the probability of generating specific egocentric cues C from input X , and $P(O|do(C))$ denotes the effect of those cues on output O .

Interventional Cue Generation $P(C|do(X))$. To improve the egocentric reasoning ability of LVLMs, given the input X including textual prompts T and images I , we introduce the observation by constructing a hierarchical set of multi-modal egocentric cues C . Specifically, for visual cues, we utilize the grounding model (Liu et al., 2023c) to generate bounding boxes, b_1 and b_2 , corresponding to each hand. When no hand is detected, the middle-lower area is designated as the cropped region. Based on the boxes coordinates, we extract a cropped region, H_r^0 , containing the hands and objects involved in the interaction, which can be formulated as:

$$H_r^0 = f(I, b_1, b_2), \quad (2)$$

where f represents the cropping function applied to image I based on the coordinates of b_1 and b_2 .

However, focusing overly on this cropped region may exclude global-level semantic information. Therefore, we construct a hierarchical set of cropped regions $\mathbf{H}_r = \{H_r^i\}_{i=0}^3$ by expanding each region by 20% around its center. Here H_r^0 represents the minimal region focusing on the interaction, while higher values of i indicate increasingly broader regions. Based on the region set \mathbf{H}_r , LVLMs are prompted to generate a corresponding hierarchical set of captions \mathbf{H}_c as follows:

$$\mathbf{H}_c = \{H_c^i\}_{i=0}^3 = \text{LVLM}(\mathbf{H}_r), \quad (3)$$

The causal effect between input X and multi-modal egocentric cues C can be formulated as:

$$P(C|do(X)) = \frac{\sum_{i=1}^{|C|} (S(H_r^i, H_c^i) + S(H_c^i, H_r^i))}{||C||}, \quad (4)$$

where c_i denotes the i -th pair of multi-modal cues H_r^i and H_c^i , with $||C||$ representing the total number of multi-modal cue pairs. The function $S(\cdot)$ is used to compute the cosine similarity.

Output Influence via Cue Mediation $P(O|do(C))$. Based on the generated multi-modal egocentric cues C , *i.e.*, mediated observations, we guide the LVLMs to perform egocentric scene understanding. Specifically, given the egocentric cues C , we construct the final input prompt P_r after mediation, which can be expressed as follows:

$$P_r = [X, C], \quad (5)$$

The prompt details are available in *supplementary materials*. Then we query the LVLMs N times, obtaining N answers $A = \{A^k\}_{k=0}^N$ using the final prompt P_r . We estimate the probability of outputs by calculating CLIP score (Deng et al., 2024) between each answer and input image as follows:

$$P(O|do(C)) = \text{CLIP}(O, I). \quad (6)$$

3.3 Reasoning Uncertainty Modeling

While the front-door adjustment helps the model focus on relevant egocentric cues, there remains the challenge of noise in these cues, *i.e.*, not all egocentric cues generated from the proposed EFDA are reliable. To address this, we introduce an uncertainty-based mechanism to selectively filter out unreliable cues. In specific, we employ the Subjective Logic (SL) (Jsang, 2018) principle to quantify the uncertainty associated with the hierarchical textual

Methods	Object			Activity	Localization		Fore.	Reasoning		
	Exist	Attr	Afford		Loc	Spatial		Count	Compar	Situated
LLaVA-1.5-7B (2023)	33.0	47.0	54.0	35.5	35.0	49.0	27.0	20.0	47.0	37.0
+ FRUIT	61.0	75.0	61.0	58.0	79.0	61.0	39.5	31.0	47.0	52.0
LLaVA-1.5-7B-LoRA (2023)	63.0	60.0	63.0	55.0	79.0	49.0	26.0	50.0	63.0	61.0
+ FRUIT	85.0	71.0	66.0	77.0	85.0	63.0	58.0	53.0	70.0	70.5
InstructBLIP-7B (2023)	50.0	33.0	45.0	47.5	77.0	38.0	40.5	18.0	43.0	67.0
+ FRUIT	61.0	57.0	49.0	50.0	90.0	57.0	45.5	31.0	56.0	70.0
Otter-I-7B (2023)	48.0	56.0	39.0	44.0	60.0	44.0	38.0	39.0	48.0	42.0
+ FRUIT	57.0	62.0	45.0	44.5	63.0	49.0	45.5	42.0	50.0	44.0
PandaGPT-7B (2023)	40.0	56.0	41.0	37.0	61.0	52.0	43.0	19.0	52.0	44.0
+ FRUIT	49.0	57.0	46.0	48.0	63.0	57.0	48.0	30.0	54.0	48.0
MiniGPT4-7B (2024)	50.0	56.0	37.0	39.0	55.0	49.0	41.5	14.0	48.0	31.0
+ FRUIT	55.0	57.0	40.0	50.0	56.0	53.0	43.0	18.0	48.0	46.0
ShareGPT4V-7B (2024)	67.0	75.0	53.0	55.5	77.0	62.0	47.0	30.0	38.0	66.0
+ FRUIT	75.0	76.0	58.0	60.5	81.0	63.0	51.0	39.0	44.0	67.0
mPLUG-Owl2-7B (2024)	58.0	61.0	44.0	41.0	84.0	45.0	36.0	44.0	54.0	45.0
+ FRUIT	67.0	74.0	52.0	57.0	86.0	48.0	56.0	50.0	62.0	56.0
Janus-Pro-7B (2025)	71.0	76.0	54.0	66.5	88.0	71.0	55.0	56.0	68.0	56.5
+ FRUIT	81.0	79.0	70.0	78.5	95.0	75.0	58.0	64.0	76.0	63.0

Table 1: Comparisons with existing 7B-sized LVLMs on the EgoThink benchmark (Cheng et al., 2024), including Object, Activity, Localization, Reasoning, and Forecasting. Results presented under the shaded area indicate they have been pre-trained on first-person data. Bolded values represent the best performance.

egocentric cues, which can detect unreliable captions within the set \mathbf{H}_c . Initially, we calculate the similarity $S(\mathbf{H}_c, \mathbf{H}_r)$ between the generated captions \mathbf{H}_c and corresponding regions \mathbf{H}_r to derive the caption evidence $\mathbf{e} = \{e_i\}_{i=0}^3$ as follows:

$$\mathbf{e} = f_e(S(\mathbf{H}_c, \mathbf{H}_r)) = \exp^{F(S(\mathbf{H}_c, \mathbf{H}_r))}, \quad (7)$$

where F denotes the ReLU activation function.

Based on the caption evidence e_i , we compute α_i and characterize the uncertainty \mathbf{U} as follows:

$$\alpha_i = e_i + 1, \quad \mathbf{U} = \frac{N}{\mathbf{S}_d}, \quad (8)$$

where $\mathbf{S}_d = \sum_{i=1}^N \alpha_i$ serves as the intensity parameter for the Dirichlet distribution. The distribution can be formulated as:

$$D(\mathbf{p}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^N p_j^{\alpha_j-1} & \mathbf{p} \in \mathcal{S}_N, \\ 0 & \text{else,} \end{cases} \quad (9)$$

where $B(\boldsymbol{\alpha})$ denotes the N -dimensional beta function, and \mathcal{S}_N represents the N -dimensional unit simplex. The uncertainty $u_i \in \mathbf{U}$ acts as a quantitative measure of the reliability of each textual cue c_i . To filter out the unreliable textual cues, we obtain the refined multi-modal cues C by setting a threshold λ , removing the noisy cues when $u_i > \lambda$. **Final Outputs.** Based on the above uncertainty-based refinement, the final answer is obtained by:

$$\max_{C, X} P(O|do(X)) = P(O|do(C))P(C|do(X)), \quad (10)$$

We choose the answer with the largest probability as the final output of LVLMs.

4 Experiments

4.1 Experimental Setup

Datasets. The EgoThink dataset (Cheng et al., 2024) consists of 700 first-person images from the Ego4D (Grauman et al., 2022) video resource, each paired with a detailed question-answer set. It is organized into six core abilities, including Activity, Localization, Reasoning, Forecasting, and Planning, which are employed to evaluate the capacity of LVLMs to understand first-person images.

Evaluation Metrics. Following (Cheng et al., 2024), we use GPT-4 (Achiam et al., 2023) as an automatic evaluator to assess generated answers. By comparing the model’s output with the reference answer, GPT-4 determines accuracy based on answer similarity. A scoring system is applied: 0 points for incorrect answers, 0.5 for partial correctness, and 1 for correct answers.

Implementation Details. In this study, we employ the PyTorch framework to implement our model. We conducted experiments utilizing 14 commonly used large vision-language models (LVLMs). To ensure a fair comparison and account for potential parameter influences, the models were divided into two groups: 7B and 13B. All models were evaluated under a zero-shot setting using the EgoThink benchmark. Given the multi-modal inputs, we first utilize the GroundingDINO (Liu et al., 2023c) identify and crop regions proximal to the hand in the images. Base on these, we generate the corresponding captions and then select appropriate image-caption pairs as hierarchical multi-modal egocentric cues.

Methods	Object			Activity	Localization		Fore.	Reasoning		
	Exist	Attr	Afford		Loc	Spatial		Count	Compar	Situated
LLaVA-1.5-13B (2023)	54.0	62.0	52.0	46.0	53.0	46.0	44.0	26.0	44.0	29.0
+ FRUIT	75.0	70.0	54.0	61.0	77.0	69.0	48.0	38.0	57.0	62.0
MiniGPT4-13B (2024)	47.0	48.0	31.0	28.0	60.0	40.0	38.5	33.0	33.0	46.0
+ FRUIT	60.0	50.0	39.0	47.5	64.0	43.0	40.0	35.0	53.0	48.0
LLaVA-NeXT-13B (2024)	65.0	72.0	36.0	34.0	60.0	57.0	28.0	43.0	46.0	18.5
+ FRUIT	86.0	79.0	69.0	69.5	78.0	69.0	43.0	44.0	65.0	40.0

Table 2: Comparisons with existing 13B-sized LVLMs on the EgoThink (Cheng et al., 2024).

These cues are introduced to make front-door adjustments for the LVLM reasoning process.

4.2 Overall Comparison Results

LVLM baselines. We evaluate the efficacy of our FRUIT method on 11 widely used LVLMs. Overall, we divide them into two groups: (1) *7B-sized Models*: LLaVA-1.5-7B (Liu et al., 2023b), InstructBLIP-7B (Dai et al., 2023), MiniGPT4-7B (Zhu et al., 2024), Otter-I-7B (Li et al., 2023a), PandaGPT-7B (Su et al., 2023), ShareGPT4V-7B (Chen et al., 2024), mPLUG-Owl2-7B (Ye et al., 2024), Janus-Pro-7B (Chen et al., 2025). (2) *13B-sized Models*: LLaVA-1.5-13B (Liu et al., 2023b), MiniGPT4-13B (Zhu et al., 2024), LLaVA-NeXT-13B (Liu et al., 2024a). More results are provided in *supplementary materials*.

Results on the EgoThink benchmark. According to the comparison on EgoThink reported in Table 1 and Table 2, we can find that: (1) Our FRUIT method demonstrates marked improvements in accuracy across various partitions of the EgoThink dataset. These results suggest that our method enhances the model’s ability to accurately extract crucial information in first-person context. (2) Moreover, even when compared to the latest LVLM Janus-Pro-7B (Chen et al., 2025) proposed by DeepSeek or LLaVA-1.5 pretrained on egocentric data, our FRUIT method shows a substantial improvement across various metrics. We hypothesize that this performance enhancement is attributed to the proposed egocentric front-door adjustments, which incorporate mediated observations to help guide LVLMs in more effectively interpreting semantics from a first-person perspective.

Results on the Planning Task. As illustrated in Figure 4, we evaluated our FRUIT method on both the planning assistant and planning navigation tasks. The results reveal several findings: (1) Our method achieved superior results in two data categories across the six models evaluated, indicating its efficacy in extracting information from a first-person perspective. (2) Compared to the re-

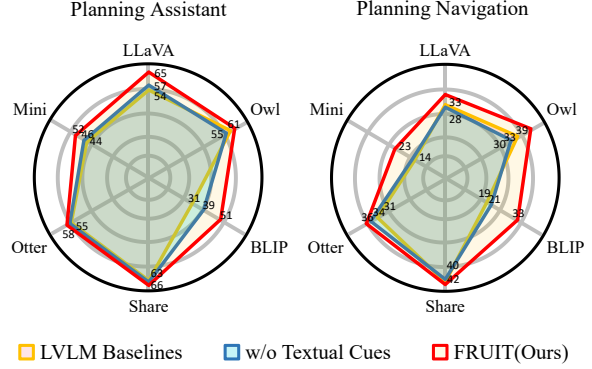


Figure 4: Performance comparison of 7B-sized LVLMs on Planning Assistant and Planning Navigation. Six models are evaluated: LLaVA-1.5 (LLaVA), mPLUG-Owl (Owl), InstructBLIP (BLIP), ShareGPT4V (Share), Otter-Image (Otter) and MiniGPT4 (Mini).

sults obtained w/o textual cues, our FRUIT method achieves superior performance across both evaluation metrics. These improvements underscore the critical role of incorporating hierarchical, multi-modal egocentric cues in capturing the key information from a first-person perspective.

4.3 Further Analysis

Ablation Study. As presented in Table 3, we list the following conclusions: (1) The comparison with No.0 and No.3 reveals that our proposed EFDA significantly enhances the performance of LVLMs on a wide range of egocentric tasks. Such results demonstrate the efficacy of integrating hierarchical multi-modal egocentric cues as mediators. It can help LVLMs capture both global-level semantics and human-object interactions, thereby facilitating reasoning from a first-person perspective. (2) From the comparison of No.1 and No.3, No.2 and No.3, we speculate that the interaction between the ICG and OICM is crucial for understanding knowledge within an egocentric context. One reasonable reason is that the ICG provides a structured representation of class hierarchies and relational mappings, while the OICM enhances this framework by adding contextual interactions that are unique to the egocentric perspective. More

No.	Components			Metrics								
	UCM	EFDA		Object			Activity	Localization		Reasoning		
		ICG	OICM	Exist	Attr	Afford		Loc	Spatial	Count	Compar	Situated
0	-	-	-	33.0	47.0	54.0	35.5	35.0	49.0	20.0	41.0	37.0
1	-	✓	-	58.0	62.0	57.5	45.5	63.0	54.5	29.0	42.0	45.5
2	-	-	✓	56.0	61.0	56.5	42.0	61.0	58.0	28.0	47.0	42.0
3	-	✓	✓	69.0	70.0	58.0	57.0	76.0	58.0	37.5	45.5	50.0
4	✓	✓	-	63.0	65.0	58.0	55.0	68.0	55.5	32.0	43.0	46.5
5	✓	✓	✓	73.0	75.0	61.0	58.0	79.0	61.0	41.0	47.0	52.0

Table 3: Ablation analysis of key components on the EgoThink benchmark dataset. The baseline is LLaVA-7B here.

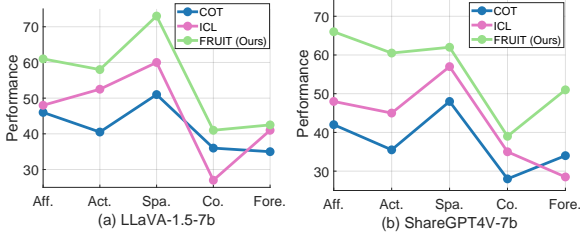


Figure 5: Performance comparison of our FRUIT method against In-Context Learning (ICL) and Chain-of-Thought (COT) prompting strategies across distinct subsets: Affordance (Aff.), Activity (Act.), Spatial (Spa.), Counting (Co.), and Forecasting (Fore.).

details are available in *supplementary materials*.

Analysis on Different Prompting Techniques.

We compare the performance of our proposed FRUIT with conventional prompting methods, including In-Context Learning (ICL) and Chain-of-Thought (CoT). Note that all methods under consideration incorporate grounded interactions to ensure a fair comparison. By observing Figure 5, we can find that: The performance comparisons reveal a substantial improvement using the FRUIT method compared to traditional prompting techniques. These results indicate that these methods fail to capture causal reasoning relationships between interacting objects and user’s intentions. In contrast, FRUIT enables LVLMs to process and reason from an egocentric perspective, reinforcing our rationale that simulating human causal reasoning can enhance egocentric reasoning.

Probability Estimation Selection in $P(O|do(C))$.

We investigate the influence of different probability estimations in $P(O|do(C))$ on Reasoning and Object tasks. Observations drawn from Figure 6 are as follows: (1) The majority voting strategy demonstrates a more accurate output in reasoning tasks involving high uncertainty, likely due to the aggregation of individual judgments, effectively mitigating the impact of outlier predictions. In contrast, the CLIP score exhibits a higher sensitivity to contextual variations, leading to more nuanced predictions, especially in tasks that require intricate

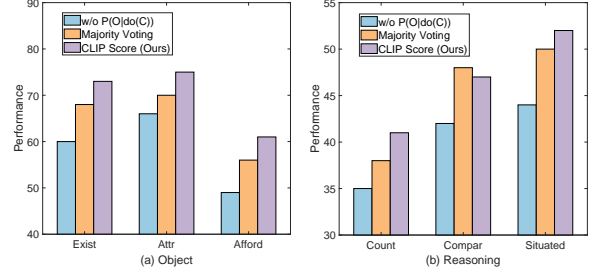


Figure 6: Effect of $P(O|do(C))$ estimation strategies. Performance compared across w/o $P(O|do(C))$, majority voting, and CLIP score (Ours).

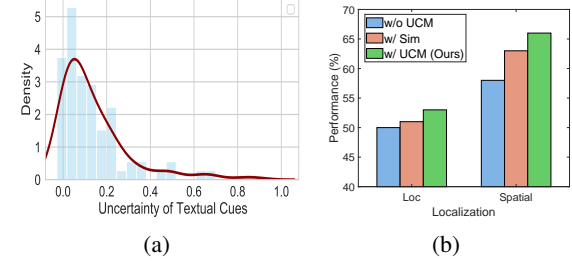


Figure 7: (a) Visualization of the uncertainty distribution of textual cues, illustrating the density of uncertainty levels. (b) Comparison on the Localization subset across different noise selection methods: without UCM, using cosine similarity (Sim), and with UCM (Ours).

understanding of semantic relationships between objects and actions. (2) When using the first predicted answer without applying $P(O|do(C))$, the performance of w/o $P(O|do(C))$ consistently suffers across all tasks. This indicates that relying solely on direct predictions, without incorporating egocentric reasoning through causal intervention, undermines the accuracy of final decision-making.

Visualization and Effect of Uncertain Cues. To further substantiate the presence of uncertain cues and validate the importance of our proposed Uncertainty Calibration Mechanism (UCM), we visualize the distribution of uncertainty for textual cues and compare the impact of various noise selection strategies, employing the MiniGPT4-7b model. As depicted in Figure 7, we can find that: (1) In Figure 7(a), the distribution of normalized uncertainty differs between regular tokens and hallucinations.

Models	P@1	P@3	P@5
LLaVA-1.5-7b (Liu et al., 2023a)	31.0	46.0	52.0
+FRUIT	59.0	79.0	80.0
ShareGPT4V-7b (Chen et al., 2024)	28.0	46.0	55.0
+FRUIT	60.0	77.0	85.0

Table 4: Evaluation of object interaction focus for LLaVA-1.5 and ShareGPT4V using our proposed P@K, with and without the FRUIT. Higher P@K means better top-K relevant object identification accuracy.

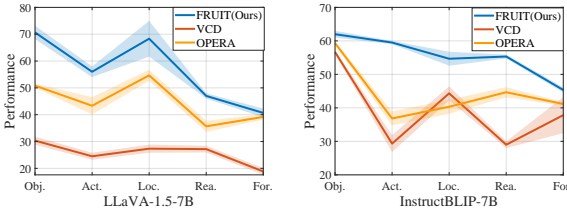


Figure 8: Performance comparison of the proposed FRUIT method against VCD (Leng et al., 2024) and OPERA (Huang et al., 2024).

Hallucinations have a higher density at the upper end of the uncertainty scale, while regular tokens cluster towards lower levels. This difference validates the effectiveness of our proposed UCM in detecting unreliable textual cues. (2) As illustrated in Figure 7(b), our FRUIT method consistently outperforms the similarity-based strategy w/ Sim. The improvement reflects the efficacy of our approach in reducing the impact of noisy cues, which in turn enhances egocentric reasoning ability.

Evaluation of Object Interaction Focus. To evaluate the LVLMS’ ability to focus on correct objects in first-person scenarios, we define the P@K, a metric that quantifies the probability that the correct answer appears in the top-K candidate predictions. For example, given 10 image-question pairs, if there are 4 correct answers within the top-5 candidate predictions from LVLMS, then $P@5 = 4/10 = 0.4$. From results in Table 6, we can list the following conclusions: (1) LVLMS with FRUIT demonstrate a higher likelihood of identifying relevant objects, due to the cues from the observation stage, which guide the model to focus on human-object interaction regions. (2) While multimodal cues from local interaction regions can generate descriptions centered on human-object interactions, they may lack essential global semantic context.

Qualitative Analysis. As shown in Figure 9, we present a qualitative comparison of the performance between the LVLMS baselines (LLaVA-1.5 (Liu et al., 2023a) and PandaGPT (Su et al., 2023)) and these models enhanced with our FRUIT method. For example, when involving multiple

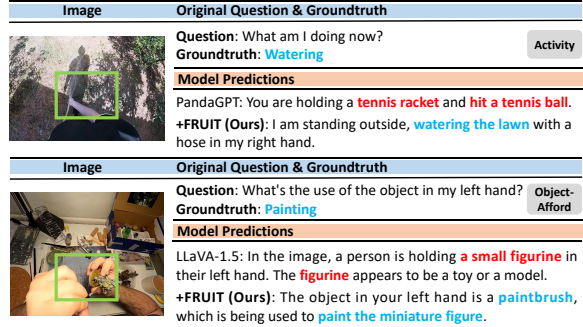


Figure 9: Comparative analysis between LVLMS baselines and our proposed FRUIT method in the Activity and Object subsets. Incorrect descriptions are highlighted in red. Correct predictions are bolded in blue.

interacting objects, LLaVA-1.5 with our proposed FRUIT can consistently maintain accurate object localization in dynamic environments. However, the LLaVA-1.5 baseline often erroneously identifies the relationships between the object and hands in egocentric context. This comparison shows the importance of our FRUIT method in focusing on human-object interaction, thus guiding LVLMS to reason from a first-person perspective.

Comparison with Hallucination Mitigation Methods. In this study, we evaluate the performance of our FRUIT method by comparing it with recent hallucination mitigation techniques, VCD (Leng et al., 2024) and OPERA (Huang et al., 2023a). As illustrated in Figure 8, our FRUIT method consistently outperforms both VCD and OPERA across all evaluation metrics. This improvement indicates that while VCD and OPERA effectively address hallucination issues in traditional third-person data, they do not enhance the reasoning capabilities of large multimodal models in first-person contexts. In contrast, our FRUIT approach directs Large Vision-Language Models (LVLMS) to perform first-person reasoning in a manner that aligns with human cognitive processes.

5 Conclusion

In this work, we introduced the Front-Door Adjustments with Uncertainty Calibration (FRUIT) framework to enhance the egocentric reasoning capabilities of large vision-language models (LVLMS). By incorporating Egocentric Front-Door Adjustments, FRUIT guide the model focus on interaction objects, while the Uncertainty Calibration Mechanism filters out unreliable information. In future work, we will explore additional strategies to further enhance LVLMS performance in first-person contexts.

6 Limitations

The current evaluation relies exclusively on image-text pairs from EgoThink, which cannot fully capture the temporal dynamics of real-world egocentric perception. Video understanding requires modeling continuous viewpoint changes and action-state transitions—capabilities our hierarchical cues currently lack due to fixed spatial grounding. Future work should extend the causal mediation framework to incorporate temporal attention mechanisms for long-horizon reasoning.

Acknowledgment

This work was supported in part by Meituan, and National Natural Science Foundation of China under Grants, China (No.62476201 and 62222203).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. 5, 11
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*. 2
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint*. 1
- Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. *European Conference on Computer Vision*. 2, 6, 8
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*. 6
- Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2024. Ego-think: Evaluating first-person perspective thinking capability of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302. 1, 2, 5, 6, 11, 12, 13, 18
- Andy Clark. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204. 2
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500. 6, 13
- Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *CoRR*, abs/2402.15300. 4
- Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR. 3
- Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012. 5
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023a. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CVPR*. 8
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427. 8
- Audun Jsang. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated. 3, 4
- Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30. 3
- Anna Kukleva, Fadime Sener, Edoardo Remelli, Bugra Tekin, Eric Sauser, Bernt Schiele, and Shugao Ma. 2024. X-MIC: cross-modal instance conditioning for egocentric action generalization. In *CVPR*, pages 26354–26363. 1
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *CVPR*. 8
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*. 6, 12, 13

- Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. 2022. Bridge-prompt: Towards ordinal action understanding in instructional videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19848–19857. 1
- Shenshen Li, Chen He, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. 2024a. Adaptive uncertainty-based learning for text-based person retrieval. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pages 3172–3180. 1
- Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Xin Liu, and Heng Tao Shen. 2024b. Multi-grained attention network with mutual exclusion for composed query-based image retrieval. *IEEE TCSVT*, 34(4):2959–2972. 2
- Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Zhe Sun, and Andrzej Cichocki. 2024c. Cross-modal attention preservation with self-contrastive learning for composed query-based image retrieval. *ACM Trans. Multim. Comput. Commun. Appl.*, 20(6):165:1–165:22. 2
- Shenshen Li, Xing Xu, Yang Yang, Fumin Shen, Yijun Mo, Yujie Li, and Heng Tao Shen. 2023b. DCEL: deep cross-modal evidential learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6292–6300. ACM. 1
- Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. 2024. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *CoRR*, abs/2403.07376. 1, 2
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744. 8
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. *Llava-next: Improved reasoning, ocr, and world knowledge*. 6, 11, 12
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. 6, 12
- Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Lily Lee, Kaichen Zhou, Pengju An, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. 2024b. Robomamba: Multimodal state space model for efficient robot reasoning and manipulation. *CoRR*, abs/2406.04339. 1
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*. 2, 4, 5
- Raanan Y. Rohekar, Yaniv Gurwicz, and Shami Nisimov. 2023. Causal interpretation of self-attention in pre-trained transformers. In *NeurIPS*. 2, 3
- Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 2012. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*. 2
- Ram Shanmugam. 2001. Causality: Models, reasoning, and inference. *Neurocomputing*, 41(1-4):189–190. 2, 3, 4
- Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*. 2, 6, 8, 11
- Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *MultiMedia Modeling*, volume 14557, pages 32–45. 2
- Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. 2023. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 3284–3294. 1
- Ronald R Yager and Liping Liu. 2008. *Classic works of the Dempster-Shafer theory of belief functions*, volume 219. Springer. 3
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. *mplug-owl: Modularization empowers large language models with multimodality*. *Preprint*, arXiv:2304.14178. 11, 13
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. *mplug-owl2: Revolutionizing multimodal large language model with modality collaboration*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051. 6, 12
- Congzhi Zhang, Linhai Zhang, and Deyu Zhou. 2024a. Causal walk: Debiasing multi-hop fact verification with front-door adjustment. In *AAAI*, pages 19533–19541. 2
- Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakhiah, Qiaozi Gao, and Joyce Chai. 2024b. Groundhog grounding large language models to holistic segmentation. In *CVPR*, pages 14227–14238. 1
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. *Minigt-4: Enhancing vision-language understanding with advanced large language models*. In *ICLR*. 1, 6, 11, 12, 13

A Appendix

A.1 Overview

In this supplementary material, we provide more analyses of our proposed FRUIT method, which are difficult to describe in the main paper due to space limitations. The content of the additional material is shown below:

- We augmented the experimental results from two state-of-the-art models using the EgoThink (Cheng et al., 2024).
- We reported experimental outcomes for two additional excellent models, evaluated using our proposed P@K metric.
- We further explore the effect of varying hyper-parameter λ .
- We conducted additional experiments to analyze the optimal number of visual egocentric cues. Specifically, we evaluated the effectiveness of the FRUIT method on images that do not contain hands.
- We included further examples of experimental results based on the EgoThink (Cheng et al., 2024).
- We provided a comprehensive list of all prompts utilized in the experiment.

A.2 Results on the EgoThink benchmark

To further validate the efficacy of our FRUIT method in enhancing the interpretation of egocentric images, we conducted experiments on the EgoThink (Cheng et al., 2024) dataset using three state-of-the-art models, mPLUG-Owl (Ye et al., 2023), LLaVA-NeXT (Liu et al., 2024a) and PandaGPT (Su et al., 2023). The results of these experiments are summarized in Table 5.

Our evaluation covered 10 categories within EgoThink (Cheng et al., 2024), excluding the planning category. Notably, our method consistently improved model performance across all 10 categories. These findings suggest that our approach significantly enhances the model’s ability to interpret first-person perspective images, including various aspects such as object identification, spatial context, and behavioral understanding.

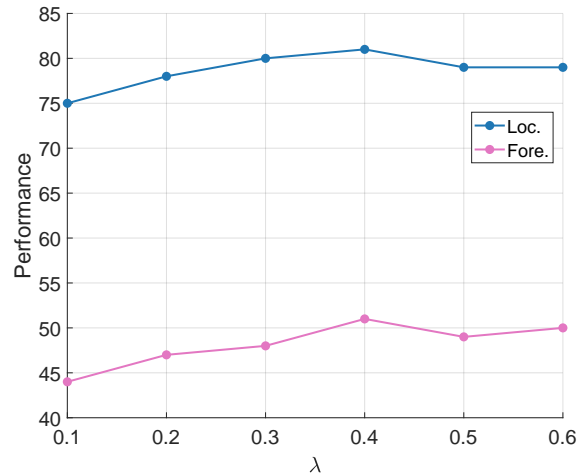


Figure 10: Comparative analysis of Localization (Loc.) and Forecasting (Fore.) performance metrics across varying values of the parameter λ within the range [0.1, 0.6].

A.3 Results on the P@K benchmark

Our P@K metric employs 100 labeled images, each representing an object with the highest degree of interaction with the character. By requiring the model to identify the top K objects with the highest human interaction and utilizing GPT-4 (Achiam et al., 2023) to verify the presence of labeled object among the K selected objects, we evaluate the model’s capability to interpret egocentric images from a first-person perspective. We implemented this metric using the two latest state-of-the-art models, LLaVA-NeXT (Liu et al., 2024a) and MiniGPT4 (Zhu et al., 2024). The experimental results in Table 6 demonstrate that our method significantly enhances the performance of the model, particularly in identifying key items. This finding substantiates the fact that our approach effectively improves the model’s comprehension of first-person images.

A.4 Effect of Varying Hyper-parameter λ

In this section, we further explore the influence of the hyper-parameter λ , which is designed to filter out unreliable cues when the uncertainty exceeds a threshold, with $\lambda = 0.4$ serving as a critical value. Figure 10 presents the performance trends for Localization (Loc.) and Forecasting (Fore.) metrics as λ varies from 0.1 to 0.6. Localization performance peaks around $\lambda = 0.4$, demonstrating stability and robustness within the range $\lambda \in [0.2, 0.5]$. This suggests that the model effectively handles moderate variations in λ . The Fore. performance

Methods	Object			Activity	Localization		Fore.	Reasoning		
	Exist	Attr	Afford		Loc	Spatial		Count	Compar	Situated
mPLUG-owl-7B (2023)	56.0	58.0	47.0	53.0	60.0	53.0	49.5	25.0	49.0	44.0
+ FRUIT	57.0	60.0	51.0	62.5	62.0	55.0	60.0	40.0	49.0	51.0
LLaVA-NeXT-7B (2024)	53.0	56.0	33.0	24.5	51.0	53.0	34.0	27.0	10.0	23.0
+ FRUIT	68.0	66.0	43.0	51.5	66.0	59.0	38.0	38.0	22.0	30.0
PandaGPT-13B (2023)	35.0	52.0	41.0	40.5	68.0	31.0	45.5	32.0	40.0	47.0
+ FRUIT	61.0	71.0	57.0	57.0	71.0	48.0	55.0	34.0	43.0	49.0

Table 5: Comparisons with existing three LVLMs on the EgoThink (Cheng et al., 2024). Bolded values represent the best performance.

Models	P@1	P@3	P@5
LLaVA-NeXT (Liu et al., 2024a)	31.0	52.0	56.0
+FRUIT	35.0	67.0	70.0
MiniGPT4 (Zhu et al., 2024)	35.0	38.0	39.0
+FRUIT	60.0	65.0	66.0

Table 6: Evaluation of object interaction focus using the proposed P@K metrics for LLaVA-NeXT and MiniGPT4, with and without the FRUIT method. Higher P@K values signify improved accuracy in identifying relevant objects across the top-K predictions.

increases slightly at intermediate λ values, achieving marginal improvement around $\lambda = 0.3$. These findings highlight the pivotal role of λ in mitigating the impact of noisy egocentric cues, with $\lambda = 0.4$ emerging as the optimal trade-off point for achieving balanced performance across both metrics.

A.5 Effect of number of visual egocentric cues

In this section, we investigate the impact of the number of images used in visual egocentric cues on model performance. Given an input prompt X including an image I and text T , we first employ a grounding model to derive hierarchical visual egocentric cues \mathbf{H}_r . To determine the optimal number of images for generating visual egocentric cues \mathbf{H}_r , we conduct experiments using 1 to 4 cropped versions of the initial images on the LLaVA-1.5-7B model (Liu et al., 2023b). The results, illustrated in Figure 11, demonstrate that using 3 cropped images yields the best performance on the EgoThink benchmark. This improvement can be attributed to the fact that a greater number of images provide richer visual egocentric information to the LVLMs, as they encompass a wider variety of focus areas at different scales. However, an excessive number of images may lead to minimal differences between them, potentially resulting in the selection of less accurate visual cues.

A.6 Results on Images without Hands

To isolate the influence of hand regions in visual egocentric cues, we selected 43 image-annotation

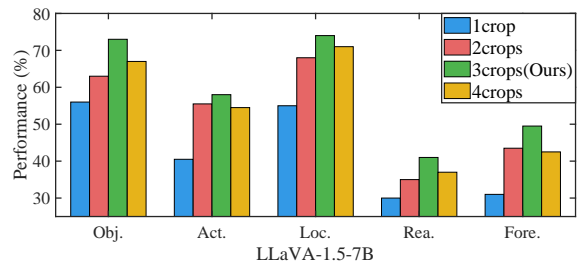


Figure 11: Comparative analysis between different number of visual egocentric cues. We use 3 crops of initial image in our FRUIT method.

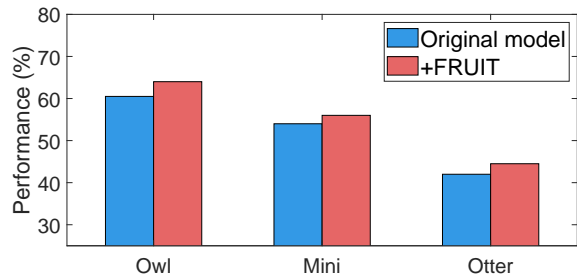


Figure 12: Comparative analysis using images that do not contain hands. Based on three LVLMs: Owl (mPLUG-owl2-7B (Ye et al., 2024)), Mini (MiniGPT4-7B (Zhu et al., 2024)) and Otter (Otter-Image-7B (Li et al., 2023a)).

pairs from the EgoThink database to evaluate the performance of our FRUIT method. The results, depicted in Figure 12, demonstrate that the FRUIT method significantly enhances the ability of LVLMs to interpret egocentric images more accurately, even in the absence of hands.

A.7 Supplementary Examples

In Figure 13, Figure 14 and Figure 15 we present a series of exemplar cases from the EgoThink (Cheng et al., 2024) dataset to demonstrate the efficacy of our FRUIT method in enhancing the accuracy of egocentric image interpretation. In each of the four instances, our FRUIT method consistently enabled the model to produce correct responses. Notably, our approach significantly enhances the model’s capacity to discern both object positions and its own spatial orientation within self-centered images,

thereby improving its comprehension from a first-person perspective. For instance, following the integration of our method, Otter-Image (Li et al., 2023a) can accurately predict subsequent actions, mPLUG-owl (Ye et al., 2023) can discern the relative position of the cutting board and the individual, MiniGPT-4 (Zhu et al., 2024) can compare items based on their proximity to the observer, and InstructBLIP (Dai et al., 2023) can identify the attributes of specific objects.

A.8 Supplement of Prompts

Follow the setting of (Cheng et al., 2024), the structure of prompt in our experiment is in Table 7. The detailed prompts are in Table 8, categorized by model type. The prompts for the planning experiments are shown in Table 9. Overall, we adopted a first-person perspective for the models and instructed them to deliver concise responses.

We conducted comparative experiments between In-Context Learning and Chain of Thought by modifying the prompt words. The prompts used are presented in Table 10. These prompts were added following the first-person perspective setting prompts provided by EgoThink(Cheng et al., 2024).

For our proposed metric P@K, we use the prompt words in Table 10 to prompt the models to generate K candidate words. When applying the FRUIT method, we introduce additional text clues with the prompt “*Here is the caption of the image: {caption}*”.

We evaluated the experimental results using GPT-4. Employing the prompts in Table 11, we instructed GPT-4 to assess the similarity between the answers generated by the model and the standard answers, and based on this assessment, determine the correctness of the answers.

Structure of prompt
[System] You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions.
[Images] Here are two images, an original image and a focused area. <image>
[Captions] Here is the caption of the human-object interaction area: {caption}
[Question] {question}

Table 7: Structure of our prompt used in EgoThink benchmark (Cheng et al., 2024).



Image	Original Question & Answer
	<p>Question: What will I do?</p> <p>Ground truth: Open the cabinet.</p> <p>Forecast</p> <p>Model Predictions</p> <p>Otter-Image: It appears that you will enter the kitchen.</p> <p>+FRUIT(Ours): I will open the cabinet door.</p>
Image	Original Question & Answer
	<p>Question: Which one is closer to me? The sink or the rubbish can on the ground?</p> <p>Ground truth: The sink.</p> <p>Reasoning Comparing</p> <p>Model Predictions</p> <p>MiniGPT4 : Rubbish can on the ground.</p> <p>+FRUIT(Ours): The sink is closer to me.</p>

Figure 13: Comparative analysis between LVLM baselines and our proposed FRUIT method in the Forecast and Reasoning Comparing. Incorrect descriptions are highlighted in red. Correct predictions are bolded in blue.


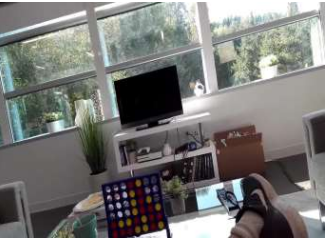
Image	Original Question & Answer
	<p>Question: Where is the cutting board, on my left or on my right?</p> <p>Ground truth: On my right.</p> <p>Localization Spatial</p> <p>Model Predictions</p> <p>mPLUG-owl: The cutting board is on my left.</p> <p>+FRUIT(Ours): On my right.</p>
Image	Original Question & Answer
	<p>Question: Am I in the bedroom?</p> <p>Ground truth: No.</p> <p>Localization Location</p> <p>Model Predictions</p> <p>PandaGPT: Yes, you are in the bedroom.</p> <p>+FRUIT(Ours): No, you are in the living room.</p>

Figure 14: Comparative analysis between LVLM baselines and our proposed FRUIT method in the Localization Spatial and Localization Location. Incorrect descriptions are highlighted in red. Correct predictions are bolded in blue.

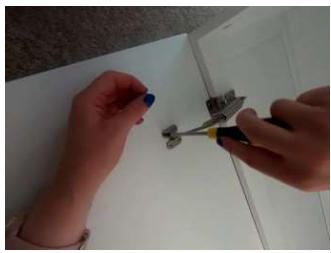

Image	Original Question & Answer
	<p>Question: What am I holding in my hand?</p> <p>Ground truth: Screwdriver.</p> <p>Object Existence</p> <hr/> <p>Model Predictions</p> <p>ShareGPT4V : In your hand, you are holding a pair of scissors.</p> <p>+FRUIT(Ours): You are holding a screwdriver in your hand.</p>
Image	Original Question & Answer
	<p>Question: Is the object I'm holding in my left hand made of transparent material or opaque material?</p> <p>Ground truth: It is made of transparent material.</p> <p>Object Attribute</p> <hr/> <p>Model Predictions</p> <p>InstructBLIP: Opaque.</p> <p>+FRUIT(Ours): The object being held in the left hand is made of transparent material.</p>

Figure 15: Comparative analysis between LVLM baselines and our proposed FRUIT method in the Object Existence and Object Attribute. Incorrect descriptions are highlighted in red. Correct predictions are bolded in blue.

Model	General Prompts
LLaVA series models	You are a person in the situation shown in the image.\n You are able to understand the visual content, \n You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns.\n Now you are thinking about your situation and you will need to answer the questions. Answer the questions in the first-person perspective.\n Keep your answer as short as possible! Keep your answer as short as possible! Keep your answer as short as possible! USER: Here are two images, an original image and a focused area. {image} Here is the caption of the human-object interaction area: {caption} Question: {question} ASSISTANT:
InstructBLIP	Please answer the following question in a few words as short as possible. Question: {question} Here is the caption of the image:{caption} Answer:
mPLUG-owl	You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions. Answer the questions in a first person perspective. Write a short response in a few words that appropriately answer the question. Keep your answer as short as possible. Here are two images, an original image and a focused area. \n <image>\n Here is the caption of the human-object interaction area: {caption} Question: {question} Short answer:
PandaGPT	Answer the following question as short as possible with a few words. \n Here is the caption of the image: {caption} \n Question: {question} \n Short Answer:
MiniGPT-4	You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions. Answer the questions in the first-person perspective. Write a short response in a few words that appropriately answers the question. End your answer with a new line. Keep your answer as short as possible in a few words! Keep your answer as short as possible! Here is the caption of the image: {caption} Question: {question} Short answer:

Table 8: Inference prompts utilized in the majority of model capabilities, except for planning.

Model	Prompts for Planning
LLaVA series models	You are a person in the situation shown in the image. \n You are able to understand the visual content, \n You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns.\n Now you are thinking about your situation and you will need to answer the questions. Answer the questions in a detailed and helpful way. USER: Here are two images, an original image and a focused area. {image} Here is the caption of the human-object interaction area: {caption} Question: {question} ASSISTANT:
InstructBLIP	Please answer the following question in a detailed and helpful way. List steps to follow if needed. Question: {question} Here is the caption of the image:{caption} Answer:
mPLUG-owl	You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions. Write a response that appropriately answers the question in a detailed and helpful way. \n Here are two images, an original image and a focused area. \n <image>\n Here is the caption of the human-object interaction area: {caption} Question: {question} Short answer:
Otter Image	You are a person in the situation shown in the image. Answer your question in a detailed and helpful way. Here is the caption of the human-object interaction area: {caption} Question: {question}
PandaGPT	Answer the following question in a detailed and helpful way.\n Here is the caption of the human-object interaction area: {caption} \n Question: {question} \n Short Answer:
MiniGPT-4	You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions. Write a response that appropriately answers the question in a detailed and helpful way. End your answer with a new line. Here is the caption of the image: {caption} Question: {question} Short answer:

Table 9: Inference prompts utilized in planning.

Method	Prompts for Analysis
In-Context Learning	When analyzing image, follow these steps step by step to solve the problem: \n 1.Understand the Question. \n 2.Observe Key Elements in the Image. \n 3.Connect Image Observations to the Question. \n 4.Formulate an Answer. \n 5.Summarize the Conclusion. \n USER: Here are two images, an original image and a focused area. {image} \n Here is the caption of the image: {caption} \n Question: {question} \n ASSISTANT:
Chain-of-Thought	Answer this question using a step-by-step Chain-of-Thought approach. Start by analyzing the question carefully and identifying the main factors involved. Then, break down the problem into smaller, manageable parts, addressing each one individually. As you progress through each step, explain your thought process and reasoning behind each conclusion you make. Continue this structured approach until you have fully explored the question, considered any potential alternatives, and arrived at a well-reasoned answer. If possible, summarize the final answer and reasoning at the end. \n USER: Here are two images, an original image and a focused area. {image} \n Here is the caption of the image: {caption} \n Question: {question} \n ASSISTANT:
P@K	You are a person in the situation shown in the image. \n You are able to understand the visual content.\n Now you are thinking about your situation and you will need to answer the question. \n USER: Which K objects am I currently interacting with? Answer the question with K objects, Use . to separate five objects! Please only answer the name of this object and do not provide any additional information! \n For example: book.phone.keyboard.bottle.paper \n ASSISTANT:

Table 10: Prompts utilized in method of in-Context learning, chain-of-Thought and P@K.

Model	Prompts for Evaluation
GPT-4(in EgoThink(Cheng et al., 2024))	[Instruction] Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider correctness and helpfulness. You will be given a reference answer and the assistant’s answer. Begin your evaluation by comparing the assistant’s answer with the reference answer. Identify and correct any mistakes. The assistant has access to an image alongwith questions but you will not be given images. Therefore, please consider only how the answer is close to the reference answer. If the assistant’s answer is not exactly same as or similar to the answer, then he must be wrong. Be as objective as possible. Discourage uninformative answers. Also, equally treat short and long answers and focus on the correctness of answers. After providing your explanation, you must rate the response with either 0, 0.5 or 1 by strictly following this format: “[rating]”, for example: “Rating: [[0.5]]”. \n [Question]\n question\n \n [The Start of Reference Answer]\n {ref answer 1} \n [The End of Reference Answer]\n \n [The Start of Assistant’s Answer]\n {answer}\n [The End of Assistant’s Answer]
GPT-4(in P@K)	There are a label and an answer. Please determine whether the objects in the answer contain the object in the label. As long as the same item is described, it is considered included. If included, answer yes. If not included, answer no. Answer with only one word!\n Label:{label}\n Answer:{answer}

Table 11: Inference prompts utilized in judgment.