EgoCross: Benchmarking Multimodal Large Language Models for Cross-Domain Egocentric Video Question Answering

Yanjun Li[1*], Yuqian Fu[2*], Tianwen Qian[1†], Qi'ao Xu[1], Silong Dai[1], Danda Pani Paudel[2], Luc Van Gool[2], Xiaoling Wang[1†]

[1]School of Computer Science and Technology, East China Normal University
[2]INSAIT, Sofia University "St. Kliment Ohridski"

Abstract

Recent advances in Multimodal Large Language Models (MLLMs) have significantly pushed the frontier of egocentric video question answering (EgocentricQA). However, existing benchmarks and studies are mainly limited to common daily activities such as cooking and cleaning. In contrast, real-world deployment inevitably encounters domain shifts, where target domains differ substantially in both visual style and semantic content. To bridge this gap, we introduce **EgoCross**, a comprehensive benchmark designed to evaluate the cross-domain generalization of MLLMs in EgocentricQA. EgoCross covers four diverse and challenging domains, including surgery, industry, extreme sports, and animal perspective, representing realistic and high-impact application scenarios. It comprises approximately 1,000 QA pairs across 798 video clips, spanning four key QA tasks: prediction, recognition, localization, and counting. Each QA pair provides both OpenQA and CloseQA formats to support fine-grained evaluation. Extensive experiments show that most existing MLLMs, whether general-purpose or egocentric-specialized, struggle to generalize to domains beyond daily life, highlighting the limitations of current models. Furthermore, we conduct several pilot studies, e.g., fine-tuning and reinforcement learning, to explore potential improvements. We hope EgoCross and our accompanying analysis will serve as a foundation for advancing domain-adaptive, robust egocentric video understanding.

Code — https://github.com/MyUniverse0726/EgoCross

1 Introduction

Egocentric videos, which capture how humans perceive and interact with the physical world from a first-person perspective, offer a rich and unique source of data for modeling human behaviors. Understanding egocentric vision is therefore highly valuable for applications such as embodied AI, wearable assistants, and human-to-robot learning. Among various egocentric tasks, video question answering (VQA) (Zhong et al. 2022) has emerged as a particularly challenging yet impactful problem. Early efforts like EgoVQA (Fan 2019), EgoTaskQA (Jia et al. 2022), and EgoSchema (Mangalam, Akshulakov, and Malik 2023) laid the groundwork for EgocentricQA by introducing dedicated benchmarks. The rapid progress of Multimodal Large Language Models (MLLMs) has further significantly advanced this field in both benchmark construction and model development. On the benchmark side, EgoThink (Cheng et al. 2024), EgoTempo (Plizzari et al. 2025), and EgoTextVQA (Zhou et al. 2025) have been proposed, targeting different aspects of the QA task. On the modeling side, a number of MLLMs specifically designed or adapted for egocentric video understanding have also emerged. Notable examples include EgoVLPv2 (Pramanick et al. 2023) and EgoGPT (Yang et al. 2025), which extend general-purpose MLLMs for EgocentricQA by training on specialized egocentric data.

Despite recent progress, most existing works remain focused on common daily-life activities, such as cooking, eating, and gardening. However, real-world applications inevitably extend beyond such scenarios. For example, in a surgical setting, a model must not only recognize a generic "cutting tool" but also precisely differentiate between instruments like a grasper, a cautery hook, and bipolar forceps. In such cases, both the visual appearance and the semantic context deviate significantly from those found in everyday activities. This naturally raises a fundamental question: *Can existing MLLMs generalize effectively to these uncommon and domain-specific scenarios?*

To answer this question, we introduce **EgoCross**, a comprehensive benchmark designed to evaluate the cross-domain generalization capabilities of MLLMs in EgocentricQA. EgoCross is built upon three core design principles: ① emphasis on cross-domain properties, ② relevance to practical applications, and ③ fine-grained, multi-dimensional model assessment. Following these principles, we carefully curated video sources and developed corresponding QA pairs to reflect real-world, high-impact use cases. Specifically, we selected surgery, industry, extreme sports, and animal perspective, as the four basic domains of our benchmark. These domains exhibit substantial visual and semantic deviations from typical daily-life scenarios, thus posing unique challenges for model generalization. Based on these video sources, we designed a structured data curation pipeline to construct QA pairs across four fundamental QA task types: *identification*, *localization*,

[*]These authors contributed equally.
[†]Corresponding authors.

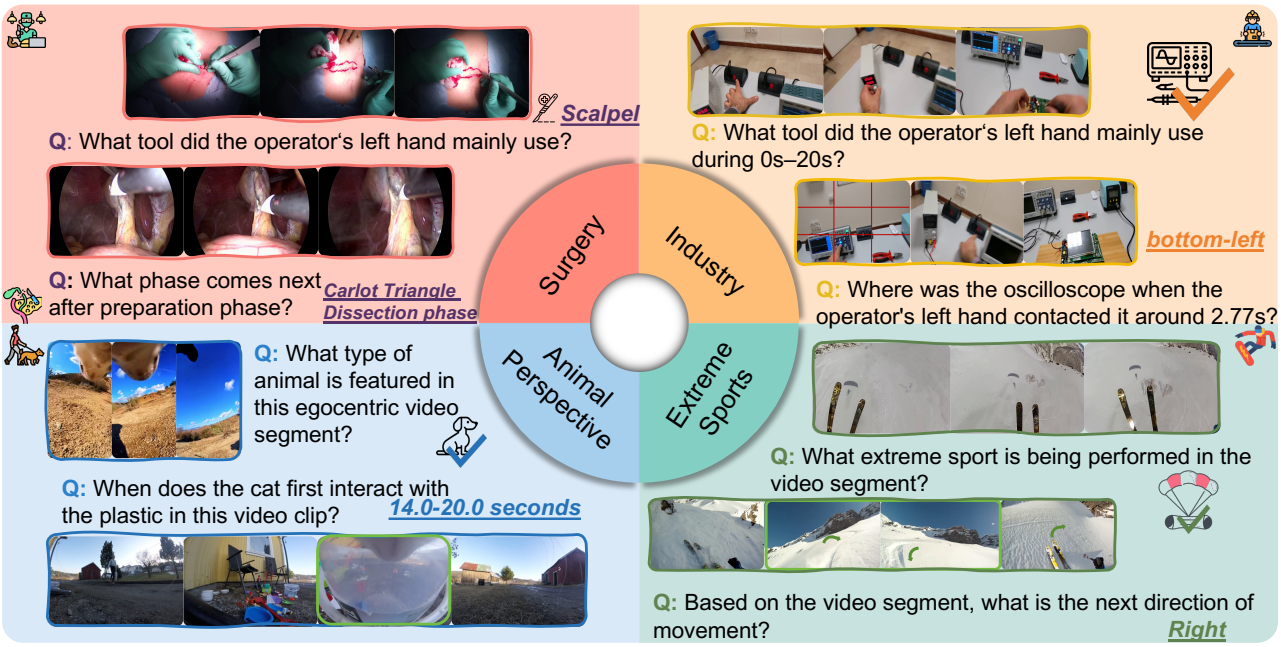


Figure 1: **Examples of Our EgoCross Benchmark.** We go beyond everyday egocentric scenarios, covering four diverse, cross-domain, application-oriented areas: Surgery, Industry, Extreme Sports, and Animal Perspective. As shown in the examples, both the visual appearances and the semantic content differ significantly from existing EgocentricQA datasets.

prediction, and *counting*, further spanning a total of 15 specific subtasks. To support both discriminative and generative evaluation protocols, each QA instance is annotated in both CloseQA (multiple-choice) and OpenQA (free-form answer) formats. In total, EgoCross consists of approximately 1,000 QA pairs across 798 egocentric video clips, forming a carefully constructed dataset that enables systematic evaluation of cross-domain generalization in EgocentricQA. A visual overview and representative examples are provided in Fig. 1.

Experiments demonstrate that most general-purpose and egocentric-specific MLLMs struggle on EgoCross, with CloseQA accuracy below 55% (random chance: 25%) and OpenQA below 35%, revealing their limitations in cross-domain settings. A notable performance drop (1.6×↓) on the same question types from EgoSchema to our EgoCross further confirms the challenge. We also explored prompt learning, fine-tuning, and reinforcement learning to assess potential improvements, offering insights for future research.

Our main contributions are summarized as follows:

- We are the first to define and motivate the task of cross-domain EgocentricQA, an underexplored yet crucial area for real-world application.
- We release EgoCross, the first cross-domain benchmark for EgocentricQA, covering four distinct domains (surgery, industry, extreme sports, and animal perspective) with ∼1k high-quality QA pairs.
- We conduct a comprehensive evaluation across 8 state-of-the-art MLLMs, quantitatively revealing their limitations beyond daily-life domains and highlighting the need for more domain-robust models.
- We provide forward-looking pilot studies, offering actionable insights and shedding light on future directions for building more generalizable and robust MLLMs.

2 Related Work

2.1 Egocentric Video Understanding

Egocentric video understanding, modeling human perception from a first-person view, has gained growing attention.Beyond basic perception (Wang et al. 2021, 2023), EgocentricQA has emerged as a key task. Initial benchmarks like EgoVQA (Fan 2019), EgoTaskQA (Jia et al. 2022), and EgoSchema (Mangalam, Akshulakov, and Malik 2023) have been joined by new datasets focusing on complex reasoning (Cheng et al. 2024), temporal understanding (Plizzari et al. 2025), and scene text (Zhou et al. 2025). The increasing data has also spurred the development of specialized models for egocentric video understanding, typically adapted from MLLMs. However, most existing work remains confined to daily-life scenarios, with limited attention paid to domain shifts. Our work fills this gap by introducing the first cross-domain testbed for EgocentricQA, emphasizing real-world, out-of-distribution targets.

2.2 MLLMs for Video Understanding

Recent MLLMs have shown remarkable capabilities in video understanding. General MLLMs such as GPT-4.1 (Achiam et al. 2023), Gemini 2.5 Pro (Comanici et al. 2025), Qwen2.5-VL (Bai et al. 2025), and InternVL (Zhu et al. 2025) achieve strong performance across a range of video tasks through extensive multimodal pretraining. In

parallel, specialized models like Video-LLaMA3 (Zhang et al. 2025) further improve temporal reasoning via dedicated architectural designs. Several MLLMs have also been tailored specifically for egocentric video understanding, including EgoVLPv2 (Pramanick et al. 2023) and EgoGPT (Yang et al. 2025). While these models perform well on third-person videos and egocentric videos from common daily scenarios, their ability to generalize to unfamiliar, domain-specific scenarios remains largely unexamined. In this work, we systematically assess how well the current state-of-the-art MLLMs generalize to cross-domain egocentric targets, revealing their limitations and offering in-depth analysis to facilitate future research in this direction.

2.3 Cross-Domain Generalization

Cross-domain generalization is a broad and long-standing challenge in computer vision. Prior work has investigated it across various tasks, including image classification (Zhu, Zhuang, and Wang 2019; Fu et al. 2023; Zhao et al. 2020), object detection (Zheng et al. 2020; Li et al. 2025; Zhang et al. 2022), and action recognition (Pan et al. 2020; Xu et al. 2022; Chen et al. 2021a), often leveraging domain transfer, data augmentation, efficient fine-tuning, and meta-learning techniques (Chen et al. 2021b). However, these efforts have primarily focused on third-person viewpoints and low-level perception tasks. In egocentric video understanding, domain shifts are particularly pronounced due to drastic variations in scenes, task semantics, and camera motion. While some works explore cross-domain few-shot recognition in egocentric videos (Hatano et al. 2024) or leverage large models for few-shot knowledge transfer (Ge et al. 2023), these remain limited to low-level perception tasks or more general settings. In contrast, EgoCross is the first benchmark specifically designed to evaluate cross-domain generalization in EgocentricQA, tackling both domain gap and high-level reasoning challenges.

3 EgoCross Dataset

In this section, we provide a comprehensive introduction to the EgoCross benchmark. We begin by discussing the selection of domains, video sources, and the taxonomy of question-answering tasks, followed by an explanation of the data curation pipeline, and conclude with it dataset statistics.

3.1 Source Selection and Task Taxonomy

Design Principles. We established key principles for domain and dataset selection, as well as question-answering task taxonomy: ① *Emphasis on Cross-Domain Properties.* We need to select domains with distinct knowledge structures, terminologies, and interactions that differ significantly from everyday scenarios, ensuring the models are challenged by unfamiliar concepts. ② *Impact on Practical Applications.* Datasets closely related to real-world applications, e.g., healthcare and industrial operations, are encouraged, as they are expected to foster progress toward practical applications of EgocentricQA. ③ *Fine-grained Multi-dimensional Model Assessment.* Tasks should span a broad range, covering diverse examination types, such as complex reasoning and spatiotemporal dependencies, and also with comprehensive evaluation metrics.

Domain and Data Source Selection. Based on the above criteria, we select four professional domains that present distinct challenges and high real-world relevance: *surgery*, *industry*, *extreme sports*, and *animal perspective*. For each, we curated one or two high-quality, open-source datasets with expert-provided meta annotations, each presenting unique perceptual, cognitive, and reasoning demands. The selected domains and the corresponding datasets are as follows:

- **Surgery.** The surgical domain represents a highly structured, knowledge-intensive scenario where precision, sequential understanding, and risk-awareness are paramount. To enrich visual diversity, we include two datasets: *EgoSurgery* (Fujii et al. 2024), which records the videos of open-heart surgeries from the surgeon's perspective, with fine-grained annotations of hand-tool interactions and surgical phases; and *CholecTrack20* (Nwoye et al. 2025), which offers laparoscopic videos of cholecystectomy procedures from a tool-centered perspective. This dataset offers a rare yet insightful egocentric perspective captured from a tool rather than a typical human operator.
- **Industry.** Complex workflows in industrial scenarios demand not only perception of fine object manipulations but also reasoning over procedural sequences and tool-usage logic. We choose *ENIGMA-51* (Ragusa et al. 2024), a dataset containing real circuit board repair tasks.
- **Extreme Sports.** Extreme sports pose unique challenges, such as rare environments, rapid camera motion, and blur, which could well test models' spatiotemporal perception and high-speed situational reasoning. We include the *ExtremeSportFPV* (Singh, Arora, and Jawahar 2017), which features first-person videos of various extreme sports, including mountain biking, skiing, and skydiving.
- **Animal Perspective.** To challenge anthropocentric bias in existing models, we introduce the animal perspective, introducing new motion patterns, camera angles, and semantic focus to the models. *EgoPet* (Bar et al. 2024), a dataset featuring egocentric views from animals such as dogs, cats, eagles, and turtles, is thus included.

The selected domains and video sources align well with our principles ① and ②.

QA Task Taxonomy. Following Principle ③, we aim to construct diverse QA pairs to comprehensively assess model capabilities. As illustrated in Fig. 3, our evaluation framework is built around four core task categories: *Identification*, *Localization*, *Prediction*, and *Counting*. Tailored to address the unique challenges of each domain, we further decompose these four broad categories into 15 specific sub-tasks, collectively forming a comprehensive evaluation framework. In the following, we provide an overview of the 4 core tasks. Detailed sub-tasks and representative examples can be found in Fig. 3 and Appendix.

- **Identification.** Identification tasks evaluate a model's ability to recognize objects, actions, and events within

a video. These tasks require domain-specific knowledge and adaptation to subtle differences in object properties or actions across contexts. For example, in surgical scenarios, instruments like forceps and scissors share similar shapes and colors, placing high demands on the models.

- **Localization.** Localization tasks assess a model's ability to identify the precise spatial or temporal location of objects, actions, or interactions. These tasks require the model to understand spatiotemporal relationships and adjust to the variations in object positioning and motion patterns across different environments. In industrial assembly, for example, locating tools or components in a cluttered workspace is particularly challenging due to partial occlusions and rapid movements of small objects.
- **Prediction.** Prediction tasks test a model's ability to forecast future actions or outcomes based on the current content. Thus, models are expected to grasp the underlying procedural or causal relationships in unseen domains. For example, in surgery, predicting the next phase of the procedure, such as transitioning from suturing to wound closure, relies on models' understanding of established surgical patterns, which can vary significantly across different domains like industrial assembly or extreme sports.
- **Counting.** Counting tasks evaluate a model's ability to track and count distinct instances or occurrences over time. These tasks require precise identification and temporal aggregation of visual elements, which becomes increasingly complex when dealing with fast-paced or dynamic scenarios. In extreme sports, for example, counting the number of tricks or jumps requires tracking high-velocity actions and differentiating between overlapping movements in a dynamic, cluttered environment.

3.2 Data Curation Pipeline

Based on the selected data sources and question categories, we developed a multi-stage curation pipeline (Fig. 2) with three key stages: meta annotation refinement, QA template design, and batch generation with quality control. The detailed procedures are described below.

Meta Annotation Refinement. Although the selected datasets provide original annotations, these are typically tailored for simpler, task-specific objectives such as 2D spatial bounding boxes for tool interactions or temporal segments for action classification. In addition, the annotation formats vary significantly between datasets. To address this, we performed a comprehensive refinement process that involved unifying annotation formats and conducting manual reviews to ensure the label accuracy. This refinement step was essential for constructing a reliable ground truth, which serves as the foundation for all subsequent QA generation.

QA Template Design. Following the task taxonomy, we manually designed 8 initial QA templates by creating two for each of the four core task categories. To enhance linguistic diversity and complexity, we employed a large language model (Gemini 2.5 pro) to expand the initial templates by generating domain-specific sub-tasks, using the original

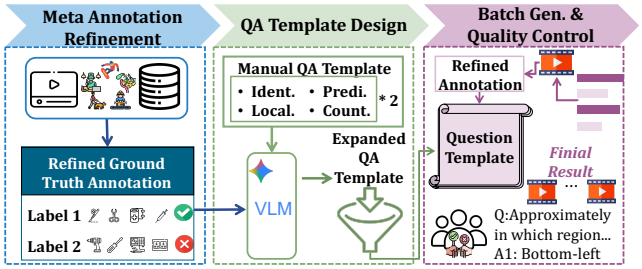


Figure 2: Data construction pipeline of EgoCross.

templates and refined annotations. All LLM-generated questions were then rigorously verified by human annotators to ensure clarity, logical consistency, and answerability based on the video content. Annotators also specified the programmatic reasoning steps and the expected answer format (e.g., object names, timestamps) for each template to ensure that the questions are both challenging and solvable.

Batch Generation and Final Quality Control. After obtaining the question templates, we perform batch instantiation to generate final QA pairs. For each sampled template, we first randomly extract a corresponding video clips based on its predifined duration, and then derive the ground-truth answer by executing the associated programmatic reasoning over the cropped clips. For comprehensive evaluation, we adopt both the traditional closed-form multiple-choice format (CloseQA) and a more flexible open-ended format (OpenQA) for the answers. In CloseQA, each question is accompanied by one correct answer and three distractors randomly sampled from the same answer type. For OpenQA, the answer consists of the full reasoning steps, which is further refined by a LLM. To ensure the data quality at scale, we conducted a final quality control check by randomly sampling and verifying 10% of QA pairs from each domain.

3.3 Dataset Statistics

Our EgoCross benchmark covers four diverse domains: Surgery, Industry, Extreme Sports (XSports), and Animal Perspective (Animal Per.), sourced from five real-world egocentric video datasets. It comprises 798 video clips and 957 QA pairs, spanning 15 sub-task types grouped into four main categories. Tab.1 summarizes key statistics of the five datasets, including the number of clips, QA pairs, and average seconds of video durations (Dur.(s)). Fig.3 further illustrates the composition of EgoCross, including: the distribution of the four primary QA task categories, the 15 sub-task types, the question counts across domains, and representative QA examples for each major capability.

4 Experiments

4.1 Experimental Setup

Evaluated Models. We select a diverse set of MLLMs spanning three categories to cover major technical paradigms: 1) To assess the current state-of-the-art performance, we include leading proprietary models: GPT-4.1 (Achiam et al. 2023) and Gemini 2.5 Pro (Comanici

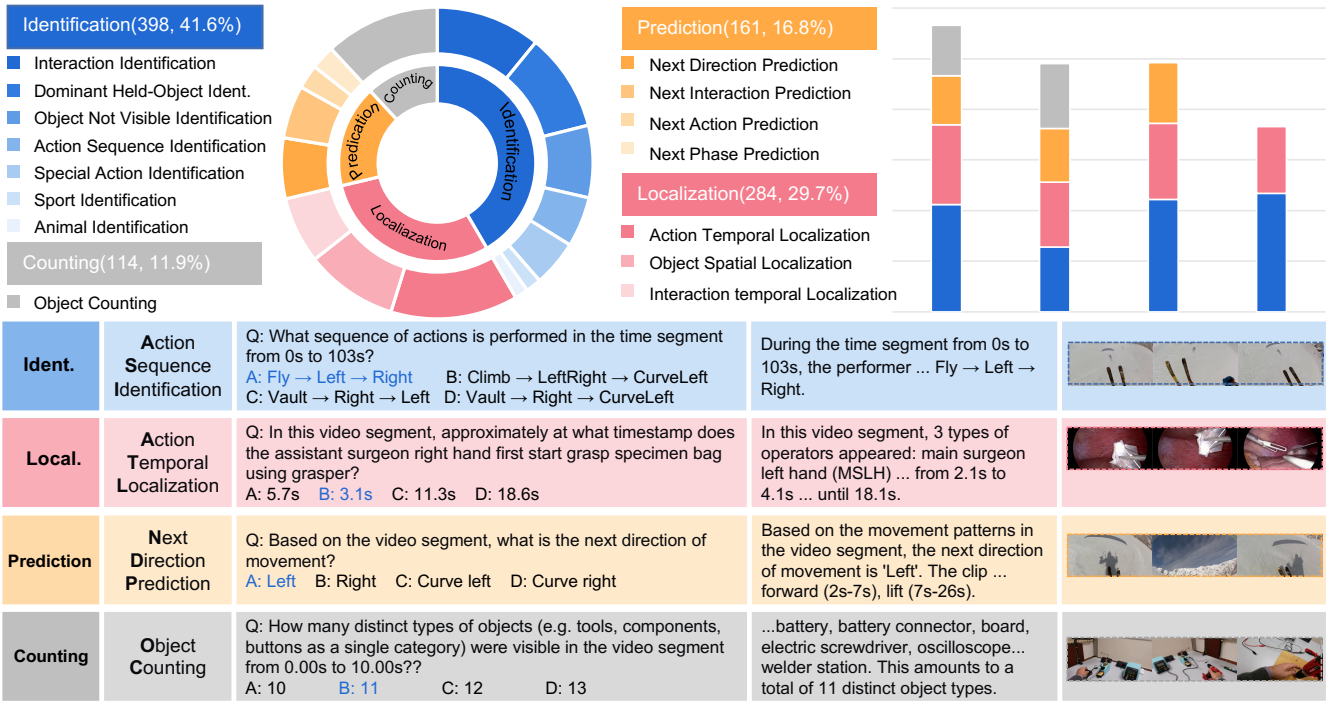


Figure 3: An overview of the EgoCross task taxonomy and statistics. (Top-left) The overall distribution of the four main task categories: Identification, Localization, Prediction, and Counting. (Top-right) The number of questions across the four primary domains. (Bottom) A selection of representative QA examples for each major capability is presented. For a more comprehensive list of examples, please see the Appendix.

Domain	Source	Clips	QA Pairs	Dur.(s)
Surgery	CholecTrack20	112	183	29.7
	EgoSurgery	100	100	20.4
Industry	ENIGMA-51	176	245	16.5
XSports	ExtremeSportFPV	242	246	13.7
Animal Per.	EgoPet	168	183	31.5
EgoCross	5 datasets	**798**	**957**	**22.5**

Table 1: Key statistics of EgoCross benchmark.

et al. 2025). 2) For open-source general-purpose MLLMs, we consider Qwen2.5-VL (3B, 7B) (Bai et al. 2025), VideoLLaMA3 (Zhang et al. 2025), and InternVL3 (Zhu et al. 2025). 3) To evaluate models tailored for egocentric understanding, we also include two egocentric-specialized models: EgoVLPv2 (Pramanick et al. 2023), and EgoGPT (Yang et al. 2025).

Evaluation Metrics. Following prior works (Fan 2019; Mangalam, Akshulakov, and Malik 2023; Plizzari et al. 2025), we use standard *accuracy* metric for CloseQA, which is calculated as the percentage of correctly answered questions. For OpenQA, we employ a two-stage evaluation process: 1) a direct exact match between the generated and ground-truth answer, and 2) if no match is found, we adopt a *LLM-as-a-Judge* approach to evaluate semantic correctness. Specifically, Qwen-MAX serves as the judge, providing a binary judgment (Correct/Incorrect) along with a detailed rationale for its decision. This ensures a robust assessment of semantic equivalence beyond simple string matching.

Implementation Details. All MLLMs are tested in a zero-shot setting with single-round inference. For video input, we extract frames at a fixed rate of 0.5 fps. For datasets that provide pre-sampled frames, we adhere to their original sampling frequency (e.g., EgoSurgery at 0.5 fps and parts of CholecTrack20 at 1 fps). No maximum frame limit is imposed to allow models to process the full temporal context. All experiments are conducted on NVIDIA A6000 GPUs. More implementation details can be found in the Appendix.

4.2 Results on EgoCross

Evaluation results are summarized in Tab. 2. We analyze the outcomes from four perspectives: 1) task-level challenges, 2) inter-domain variance, 3) model-wise performance, and 4) metric-type analysis.

Task-level Challenges. Most evaluated MLLMs struggle to perform well on our EgoCross benchmark, with average scores falling below 55% on CloseQA and below 35% on OpenQA. Considering that the random guess accuracy for CloseQA is 25%, these results suggest that the models indeed face substantial challenges in this benchmark. Furthermore, excluding the top performance from proprietary models (Gemini 2.5 Pro and GPT-4.1), the remaining

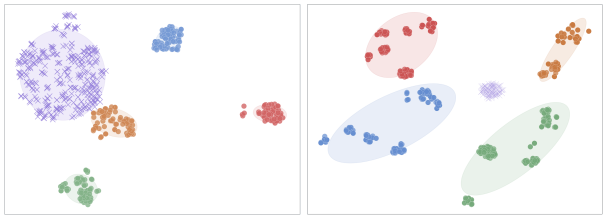
Models	Surgery		Industry		XSports		Animal Per.		Overall	
	Closed	Open	Closed	Open	Closed	Open	Closed	Open	Closed	Open
Proprietary MLLMs										
GPT-4.1	57.24	39.58	**45.71**	12.24	43.09	20.33	64.48	34.43	52.63	26.65
Gemini 2.5 Pro	**61.48**	**42.40**	37.55	**24.49**	**43.90**	**21.54**	**68.85**	**49.18**	**52.95**	**34.40**
Open-source MLLMs										
Qwen2.5-VL-3B	35.69	16.96	36.33	6.94	36.59	6.91	41.53	28.42	37.54	14.81
Qwen2.5-VL-7B	46.29	21.55	37.55	22.04	41.87	6.91	53.55	31.15	44.82	20.41
VideoLLaMA3-7B	39.22	15.90	40.82	13.47	37.80	13.41	50.27	32.24	42.03	18.76
InternVL3-8B	47.00	17.67	33.06	11.84	41.06	11.38	49.18	30.60	42.58	17.87
Egocentric MLLMs										
EgoVLPv2	26.50	-	34.69	-	23.17	-	24.04	-	27.10	-
EgoGPT	31.80	13.07	24.49	10.20	24.80	13.82	41.53	26.78	30.66	15.97

Table 2: Evaluation results of MLLMs on EgoCross. All scores are reported in percentages. The best results are marked in **bold**, and the second-best are underlined. EgoVLPv2 is not evaluated on open-set tasks due to its model architecture.

models perform notably worse, achieving less than 45% on CloseQA and under 20% on OpenQA. These observations collectively underscore both the difficulty and value of the proposed EgoCross benchmark, while also highlighting the current limitations of state-of-the-art MLLMs in handling cross-domain tasks.

Inter-Domain Variance. Across target domains, we observe varying levels of difficulty, ranging from relatively easy (Animal Perspective), middle-hard (Surgery) to particularly challenging (Extreme Sports, Industry). To further investigate inter-domain variance, we visualize t-SNE embeddings of EgoSchema and the four out-of-domain targets, using CLIP (Radford et al. 2021) as a modality-aligned feature extractor for both visual and textual representations (Fig. 4). The analysis highlights three key findings: 1) All target domains are clearly separated from EgoSchema and from each other in both visual and textual spaces, confirming the dataset's cross-domain and diverse nature. 2) Within each of our target domains, textual features form sub-clusters, indicating the richness of our QA pairs. 3) The distributions help explain domain difficulty: Animal Perspective is closest to EgoSchema in both modalities, aligning with its relatively easier performance. In contrast, Industry and Extreme Sports are farthest, consistent with their higher difficulty. Surgery seems as a visual outlier but achieves relatively strong performance, suggesting that advanced MLLMs may be more robust to perceptual variation while still challenged by deeper semantic reasoning.

Model-wise Performance. As briefly discussed above, the two proprietary MLLMs achieve the highest overall performance, with Gemini 2.5 Pro outperforming GPT-4.1. This superiority is expected, given their access to large-scale training data and advanced algorithmic design. Following them are the open-source models, including Qwen2.5-VL, VideoLLaMA3, and InternVL3. Compared to proprietary models, they exhibit clear performance drops on both CloseQA and OpenQA tasks, highlighting the substantial gap that remains in tackling cross-domain egocentric understanding within the open-source community. Surprisingly, the egocentric-specific models (EgoVLPv2, and EgoGPT) perform the worst, despite being explicitly designed and trained on egocentric video data. Their failure, in contrast to general-purpose models, more clearly underscores the challenge of cross-domain generalization.



(a) Visual features t-SNE. (b) Text features t-SNE.

Figure 4: t-SNE visualization of text and visual features. EgoCross domains are color-coded: Surgery (red), Industry (blue), ExtremeSports (green), Animal Perspective (orange) and Daily-activity (purple).

Metric-type Analysis. We further analyze the results under different evaluation metrics, namely CloseQA and OpenQA. Since CloseQA simplifies the task by providing explicit candidate answers, models naturally achieve higher accuracy in CloseQA compared to OpenQA. Additionally, we observe that CloseQA scores tend to be more stable across different MLLMs, while OpenQA is more sensitive to variations. For example, GPT-4.1 and Gemini 2.5 Pro achieve nearly identical scores on CloseQA (52.63 vs. 52.95), but differ noticeably on OpenQA (26.65 vs. 34.40). These observations confirm that OpenQA presents a more challenging setting, reflecting a common limitation of current MLLMs: their generative abilities are generally weaker than their judgment capabilities. By evaluating both OpenQA and CloseQA, we aim to comprehensively assess the generative and judgmental capacities of MLLMs in cross-domain scenarios.

In addition to dataset-level experiments and analysis, we

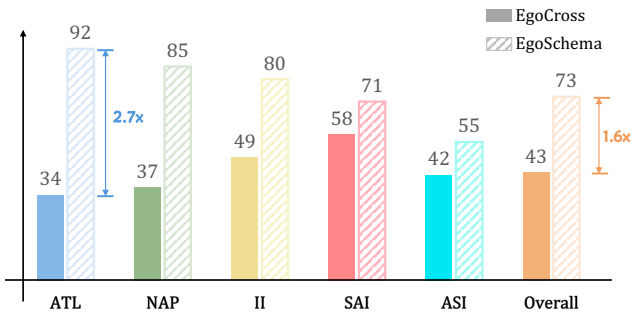


Figure 5: In-domain and cross-domain accuracy comparison across five QA types: Action Temporal Localization (ATL); Next Action Prediction (NAP); Interaction Identification (II); Special Action Identification (SAI); Action Sequence Identification (ASI). The results highlight the performance gap when evaluating on novel domains.

conduct a more fine-grained evaluation across different QA types. Due to space limitations, detailed results and discussions are provided in the Appendix.

4.3 More Analysis on Cross-Domain Gap

In Sec. 4.2, we demonstrate that domain gaps significantly contribute to the overall low performance. To further investigate this effect and highlight its unique presence in EgoCross, we compare model performance between our benchmark and EgoSchema (Mangalam, Akshulakov, and Malik 2023), a typical daily-life egocentric dataset featuring common activities like cooking and cleaning. To ensure a comparison across analogous QA types, we devise a semi-automated QA type categorization process to align EgoSchema's QA pairs with our tasks. The detailed procedure is provided in the Appendix. Based on these aligned QA pairs, we evaluate Qwen2.5-VL, which achieves the best results among open-source MLLMs, on tasks from both EgoSchema (in-domain) and EgoCross (cross-domain) using the CloseQA protocol.

Results in Fig. 5 reveal a consistent and significant performance drop across all comparable QA types. For instance, performance on *action temporal localization* drops from an impressive 92.31% on in-domain EgoSchema to just 34.13% on the novel domains of surgery, industry, and extreme sports in EgoCross. Similarly, *next action prediction* accuracy falls from 85.71% to 37.50%. The overall accuracy also drops from 73.58% to 43.14%, quantifying the substantial penalty incurred by the domain shift. This task-level analysis reinforces our core finding: despite strong results on existing egocentric benchmarks, current MLLMs lack robustness when applied to unseen, domain-specific settings, a limitation largely overlooked in prior work.

4.4 Pilot Studies

We proactively conduct several pilot studies to explore potential solutions for improving cross-domain egocentric QA.

Method	Surgery	Industry	XSports	Animal Per.	Avg.
Baseline*	46.29	37.55	41.87	53.55	44.82
Baseline	37.35	35.71	34.72	43.40	37.80
+Prompt	44.58	34.29	52.78	43.40	43.76
+SFT	37.35	52.86	40.28	43.40	43.47
+RL	**49.40**	**61.43**	**54.17**	**75.47**	**60.12**

Table 3: CloseQA accuracy in pilot studies. "+SFT" and "+RL" denote supervised fine-tuning and reinforcement learning, respectively. * denotes the baseline without vLLM acceleration, it is marked in gray as vLLM acceleration causes slight performance degradation.

Specifically, we investigate three techniques: prompt learning, supervised fine-tuning (SFT), and reinforcement learning (RL). Since both SFT and RL require labeled data, we randomly split the initial test QA pairs into training and testing sets with a 70%:30% ratio. We adopt Qwen2.5-VL-7B as the baseline, and apply vLLM (Kwon et al. 2023) for model acceleration. CloseQA results are shown in Tab. 3.

The results provide several insights: *1) Overall Trend.* Each method, whether prompting (without labeled data) or SFT/RL (requiring labeled data), improves performance to some extent. This suggests that refining prompt designs, expanding labeled data, and advancing algorithms are all promising directions for future exploration. *2) Impact of SFT.* SFT boosts accuracy in domains like Industry (nearly 20% improvement). However, in some domains, such as Animal Per. no improvement is observed. This may stem from the inherently higher base performance in this domain, which is closer to natural domains. It could be caused by the limited number of training samples (Animal Per. has only 128 samples, far fewer than other domains.), which restricts the effectiveness of SFT. *3) Effectiveness of RL.* RL shows the most significant improvement across all domains (an average increase of 22%). We attribute this to RL's ability to learn from a broader range of interactions and feedback during training. The trial-and-error process enables the model to better handle longer sequences and more complex decision-making tasks, allowing it to dynamically adapt to the unique challenges of each domain.

5 Conclusion

In this work, we present EgoCross, a new benchmark for evaluating the cross-domain generalization ability of Multimodal Large Language Models (MLLMs) in egocentric video question answering. EgoCross comprises approximately 1k QA pairs based on video clips carefully collected and curated from four diverse and realistic domains: surgery, industry, extreme sports, and animal perspective, supporting both CloseQA and OpenQA for fine-grained evaluation. Our extensive evaluation reveals that current SOTA MLLMs struggle to generalize to these unfamiliar domains, despite strong performance on existing benchmarks. Additionally, we further explore several potential techniques to improve cross-domain generalization. We believe that EgoCross, together with our experiments and analysis, offers a valuable foundation for future research.

Acknowledgments

This work was supported by NSFC grant (No. 62136002 and 62477014), Ministry of Education Research Joint Fund Project (8091B042239), and Fundamental Research Funds for the Central Universities. Project supported by Shanghai Municipal Science and Technology Major Project (2025SHZDZX025G16).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bar, A.; Bakhtiar, A.; Tran, D.; Loquercio, A.; Rajasegaran, J.; LeCun, Y.; Globerson, A.; and Darrell, T. 2024. Egopet: Egomotion and interaction data from an animal's perspective. In *European Conference on Computer Vision*.

Chen, H.-P.; Le, T.-A.; Nguyen, T.-D.; D-Phuc, N.; Tran, M.-H.; Tran, T.-A.; Ho, T.-B.; Nguyen, A.-T.; and Nguyen, V.-K. 2021a. M3Net: Multi-view Encoding, Matching, and Fusion for Few-shot Fine-grained Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Chen, Y.; Li, R.-J.; Wang, S.-Q.; Zhang, B.-J.; Zhang, C.; and Liu, C.-L. 2021b. Blockmix: Meta regularization and self-calibrated inference for metric-based meta-learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, 6779–6787.

Cheng, S.; Guo, Z.; Wu, J.; Fang, K.; Li, P.; Liu, H.; and Liu, Y. 2024. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Fan, C. 2019. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Fu, Y.; Xie, Y.; Fu, Y.; and Jiang, Y.-G. 2023. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Fujii, R.; Hatano, M.; Saito, H.; and Kajita, H. 2024. Egosurgery-phase: a dataset of surgical phase recognition from egocentric open surgery videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Ge, Y.; Liu, Z.; Li, J.; Wang, Z.; Liu, Z.; Wu, G.; Wang, Y.; Zhu, Y.; Jin, G.; Yin, F.; and Tao, D. 2023. Connecting giants: Synergistic knowledge transfer of large multimodal models for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 36.

Hatano, M.; Hachiuma, R.; Fujii, R.; and Saito, H. 2024. Multimodal cross-domain few-shot learning for egocentric action recognition. In *European Conference on Computer Vision*.

Jia, B.; Lei, T.; Zhu, S.-C.; and Huang, S. 2022. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Li, Y.; Qiu, X.; Fu, Y.; Chen, J.; Qian, T.; Zheng, X.; Paudel, D. P.; Fu, Y.; Huang, X.; Van Gool, L.; et al. 2025. Domain-RAG: Retrieval-Guided Compositional Image Generation for Cross-Domain Few-Shot Object Detection. *arXiv preprint arXiv:2506.05872*.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*.

Nwoye, C. I.; Elgohary, K.; Srinivas, A.; Zaid, F.; Lavanchy, J. L.; and Padoy, N. 2025. CholecTrack20: A Multi-Perspective Tracking Dataset for Surgical Tools. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pan, B.; Cao, Z.; Adeli, E.; and Niebles, J. C. 2020. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI conference on artificial intelligence*.

Plizzari, C.; Tonioni, A.; Xian, Y.; Kulshrestha, A.; and Tombari, F. 2025. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Pramanick, S.; Song, Y.; Nag, S.; Lin, K. Q.; Shah, H.; Shou, M. Z.; Chellappa, R.; and Zhang, P. 2023. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*.

Ragusa, F.; Leonardi, R.; Mazzamuto, M.; Bonanno, C.; Scavo, R.; Furnari, A.; and Farinella, G. M. 2024. ENIGMA-51: Towards a Fine-Grained Understanding of Human Behavior in Industrial Scenarios. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Singh, S.; Arora, C.; and Jawahar, C. 2017. Trajectory aligned features for first person action recognition. *Pattern Recognition*.

Wang, J.; Luvizon, D.; Xu, W.; Liu, L.; Sarkar, K.; and Theobalt, C. 2023. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, X.; Zhu, L.; Wang, H.; and Yang, Y. 2021. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Xu, Y.; Cao, H.; Mao, K.; Chen, Z.; Xie, L.; and Yang, J. 2022. Aligning correlation information for domain adaptation in action recognition. *IEEE Transactions on Neural Networks and Learning Systems*.

Yang, J.; Liu, S.; Guo, H.; Dong, Y.; Zhang, X.; Zhang, S.; Wang, P.; Zhou, Z.; Xie, B.; Wang, Z.; et al. 2025. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, C.-B.; Li, G.-X.; Liu, H.-B.; Zhang, Y.-P.; and Li, X. 2022. Divide-and-conquer: Confluent triple-flow network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5): 2356–2368.

Zhao, J.-M.; Zhang, R.-Y.; Jia, Y.-J.; Zhang, Y.-W.; and Zhang, Q. 2020. Learning attention-guided pyramidal features for few-shot fine-grained recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1886–1894.

Zheng, Y.; Huang, D.; Liu, S.; and Wang, Y. 2020. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Zhong, Y.; Xiao, J.; Ji, W.; Li, Y.; Deng, W.; and Chua, T.-S. 2022. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*.

Zhou, S.; Xiao, J.; Li, Q.; Li, Y.; Yang, X.; Guo, D.; Wang, M.; Chua, T.-S.; and Yao, A. 2025. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Zhu, Y.; Zhuang, F.; and Wang, D. 2019. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proceedings of the AAAI conference on artificial intelligence*.