Exo2Ego: Exocentric Knowledge Guided MLLM for Egocentric Video Understanding

Haoyu Zhang[1,2], Qiaohui Chu[1,2], Meng Liu[3,4*], Haoxiang Shi[1,2], Yaowei Wang[1,2*], Liqiang Nie[1*]

[1]Harbin Institute of Technology (Shenzhen)
[2]Pengcheng Laboratory
[3]Shandong Jianzhu University
[4]Zhongguancun Academy
{zhang.hy.2019, qiaohuichu8599, mengliu.sdu, shihaoxiang1999, nieliqiang}@gmail.com
wangyw@pcl.ac.cn

Abstract

AI personal assistants, deployed through robots or wearables, require embodied understanding to collaborate effectively with humans. However, current Multimodal Large Language Models (MLLMs) primarily focus on third-person (exocentric) vision, overlooking the unique challenges of first-person (egocentric) videos. Additionally, high acquisition costs limit data size, impairing MLLM performance. To address these challenges, we propose learning the mapping between exocentric and egocentric domains, leveraging the extensive exocentric knowledge within existing MLLMs to enhance egocentric video understanding. To this end, we introduce Ego-ExoClip, a pre-training dataset comprising 1.1M synchronized ego-exo clip-text pairs derived from Ego-Exo4D, together with the instruction-tuning dataset EgoIT, which is collected from multiple sources to enhance the model's instruction-following capabilities. Building upon the datasets, we propose a migration strategy and further design a progressive mapping learning pipeline with three stages: Demonstrator Self-Preparation, Demonstrator-Learner Guidance, and Learner Self-Practice. Extensive experiments across diverse egocentric tasks reveal that existing MLLMs perform inadequately in egocentric video understanding, while our model significantly outperforms these leading models.

Code — https://reurl.cc/Ebpyrm

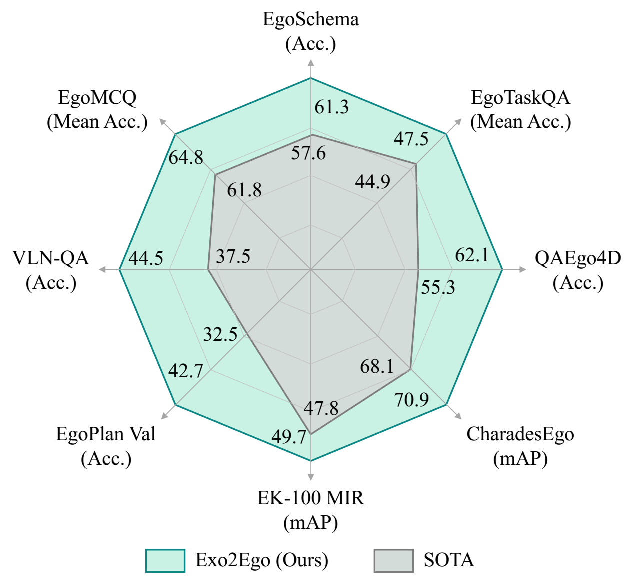

Figure 1: Our Exo2Ego model achieves optimal results across various egocentric video understanding tasks, with the detailed results presented in Table 2.

Introduction

An outsider can see things more clearly or objectively than those involved.

– The Old Book of Tang

Embodied cognition theory posits that cognitive processes are fundamentally shaped by interactions with the physical environment (Johnson 2015). In this context, understanding egocentric videos that capture human active perception and fine-grained action from a first-person perspective provides a vital pathway for artificial intelligence to understand human experiences. The challenge of analyzing egocentric videos arises from the dynamic interplay between the movements of the camera-wearer (i.e., self) and the surrounding environment (i.e., other), which distinguishes them from exocentric videos observed from a third-person perspective, where this direct experience is absent. The potential applications of this technology span a wide range, including visual aids, smart glasses, and immersive experiences in virtual and augmented reality.

In response to the increasing demand for effective egocentric video analysis, several research tasks have emerged, such as egocentric action recognition (Shiota et al. 2024; Zhang et al. 2024a), episodic memory (Feng et al. 2024; Wang et al. 2022; Feng et al. 2025), human-object interaction (Jiang et al. 2023), and action anticipation (Chu et al. 2025b,a). *However, the diverse architectures employed across these tasks lead to significant fragmentation in the*

*Corresponding authors.

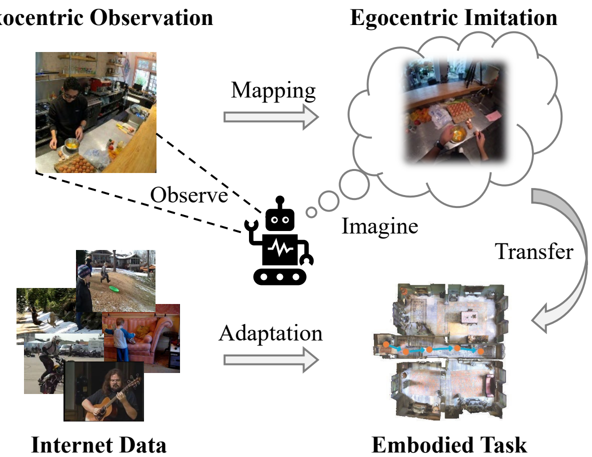

Figure 2: Understanding how humans map observed activities to their own behaviors enhances embodied cognition across various downstream tasks by transferring insights from existing Internet (i.e., exocentric) data.

field. With the success of existing Multimodal Large Language Models (MLLMs) (Wang et al. 2025; Zhang et al. 2025), there is an emerging trend toward developing a unified egocentric MLLM architecture capable of addressing multiple tasks within the egocentric domain. A significant obstacle in this direction is the limited availability of egocentric video data, which is crucial for effectively training large-capacity models (Wang, Singh, and Torresani 2023). Consequently, an increasing number of methods (Li et al. 2021; Xue and Grauman 2023; Luo et al. 2025; Dou et al. 2024) are now retrieving suitable exocentric videos from the predefined set to aid in model training for downstream tasks. While these approaches exhibit empirical advantages in transferring knowledge from exocentric to egocentric representations, *they incur additional retrieval time and suffer from misalignment, leading to instability in model performance.*

Drawing inspiration from cognitive science (Shi et al. 2024; Zhang et al. 2023), which demonstrates that children learn by observing others' behavior (i.e., an exocentric view) and mapping these observations onto their own experiences (i.e., an egocentric perspective), we conceptualize the exocentric video observer as the "demonstrator" and the egocentric video interpreter as the "learner". Our objective is to establish a robust mapping pattern between the two, as illustrated in Figure 2, offering two key advantages: 1) *Weak Dependence.* Leveraging the learned mappings, the model can flexibly tackle downstream tasks without relying on cross-domain data, thereby mitigating additional biases. 2) *Strong Generalization.* By emulating human learning, these mapping techniques significantly enhance the model's generalization capabilities, reducing the need for extensive egocentric training data in MLLMs.

To further this approach, we propose an exocentric-to-egocentric migration strategy with an egocentric self-consistency mechanism grounded in cross-view behavior invariance. By embedding this strategy within MLLMs, we design a mapping learning pipeline comprising three stages: **Demonstrator Self-Preparation**, **Demonstrator-Learner Guidance**, and **Learner Self-Practice.** By transferring exocentric knowledge, this pipeline boosts egocentric understanding and improves downstream performance. Furthermore, to satisfy the demand for synchronized cross-domain data, we introduce the **Ego-ExoClip** dataset, an advanced first- and third-person video-text dataset containing 1.1M egocentric-exocentric clip-text pairs. And we also construct an enriched instruction-tuning dataset, **EgoIT**, sourced from multiple egocentric contexts to improve the instruction-following capabilities. As shown in Figure 1, abundant results demonstrate that our framework achieves superior results, highlighting its effectiveness in advancing egocentric video understanding.

Contributions: **1**) Based on behavior invariance, we propose a knowledge transfer strategy and integrate it into the training paradigm of MLLMs, resulting in a three-stage mapping learning process. **2**) To support training, we construct a synchronized exo-ego video-text dataset, Ego-ExoClip, as well as an instruction-tuning dataset, EgoIT, collected from multiple sources. **3**) Extensive experimental results on various benchmarks validate the effectiveness of our proposed datasets and methods, and further promote research in embodied cognition.

Related Work

Egocentric Video Understanding. Egocentric videos offer a distinctive perspective for active engagement with the physical world, capturing human interactions from a first-person viewpoint. A variety of datasets have been developed to support research in this field, including EPIC-KITCHENS (Damen et al. 2020), which focuses on kitchen activities; Charades-Ego (Sigurdsson et al. 2018), which highlights diverse everyday tasks; and Ego4D (Grauman et al. 2022), which provides a globally diverse collection of egocentric videos. These datasets have driven the growth of multiple research areas, such as video-text pre-training (Lin et al. 2022; Pramanick et al. 2023), video question answering (Di and Xie 2024; Zhang et al. 2024b, 2021), human-object interaction (Akiva et al. 2023), action anticipation (Ragusa, Farinella, and Furnari 2023; Shen and Elhamifar 2024), episodic memory (Ramakrishnan, Al-Halah, and Grauman 2023; Feng et al. 2024), and pose perception (Millerdurai et al. 2024). Despite these advancements, the high cost of data collection remains a major limitation, constraining dataset scale and, in turn, the effectiveness of the models trained on them. In response, recent work (Wang et al. 2023; Xue and Grauman 2023; Truong and Luu 2024) has begun exploring the use of unpaired exocentric videos to enhance egocentric video comprehension. However, leveraging unpaired exocentric data introduces inherent biases that can negatively affect model stability and accuracy. To address this issue, we introduce a pairwise egocentric-exocentric dataset with diverse activities and leverage the exocentric knowledge embedded in MLLMs to improve egocentric video understanding.

Multimodal Large Language Model. Large Language Models (LLMs) (Touvron et al. 2023; Wen et al. 2024),

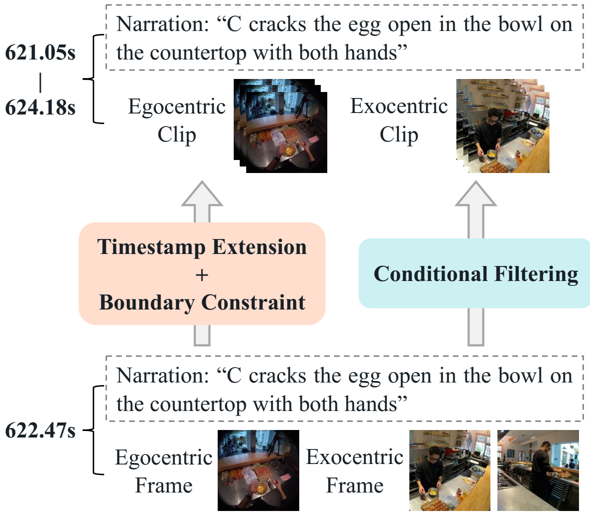

Figure 3: Illustration of the construction of Ego-ExoClip.

trained on vast datasets from the Internet, have demonstrated remarkable proficiency across a wide range of language tasks through text generation. Building upon the success of LLMs and the promise of multimodal data, there is growing scholarly interest in exploring MLLMs. Early MLLMs focus on integrating static images with text, with landmark work such as Flamingo (Alayrac et al. 2022) and BLIP-2 (Li et al. 2023), significantly expanding the capabilities of LLMs for multimodal tasks. Recent open-source MLLMs, including LLaVA (Liu et al. 2024) and InstructBLIP (Dai et al. 2023), further advance visual understanding by introducing visual instruction-tuning data. Beyond static images, recent studies have started to explore the incorporation of video data into LLMs, unlocking new potential for video comprehension tasks. Innovations such as LLaMA-VID (Li, Wang, and Jia 2023), Video-LLaVA (Lin et al. 2023), and VideoLLaMA2 (Cheng et al. 2024) aim to improve the alignment of video data with LLMs. Additionally, methods like VideoChatGPT (Maaz et al. 2023) and VideoChat2 (Li et al. 2024) utilize ChatGPT to generate video instruction-tuning data, enhancing instruction-following capabilities. Despite these advancements, existing MLLMs exhibit limited performance on egocentric video understanding tasks (Wang, Singh, and Torresani 2023). Therefore, our objective is to unify various downstream tasks and significantly enhance the capabilities of MLLMs in the embodied cognition domain.

Dataset Construction

In this section, we provide a detailed overview of the constructed Ego-ExoClip dataset, which is employed in Stages 1 and 2, along with the instruction-tuning dataset, EgoIT, developed for Stage 3 of the training process.

Ego-ExoClip

The Ego-ExoClip dataset is derived from Ego-Exo4D (Grauman et al. 2024), which comprises 5,035 grouped videos ranging in length from 1 to 42 minutes. Each group contains one egocentric video paired with 4–5 simultaneously captured exocentric videos. Most of these videos include dense timestamp-level narrations provided by two annotators, detailing the camera wearer's activities and interactions. For example, as shown in Figure 3, the narration "C cracks the egg open in the bowl on the countertop with both hands" corresponds to an event occurring at 622.47 seconds, where "C" refers to the camera-wearer.

To ensure data quality and consistency, we filter out videos without narrations, those lacking UID mappings[1], and those designated for validation or testing in benchmark challenges. This process yields a refined dataset of 2,925 video groups, totaling 15,478 videos, all validated to guarantee lossless quality. For incorporating narration diversity, we retain textual annotations from both narrators, resulting in 623.6 hours of videos and 261.3K narrations.

Given that the narrations in Ego-Exo4D are annotated at single timestamp rather than as time intervals, we extend the timestamp-level narrations to the clip-level. Specifically, narrations are organized as a sequence of sentences $\{\mathcal{S}_0, ..., \mathcal{S}_n\}$ with corresponding timestamps $\{t_0, ..., t_n\}$, indicating that an event i described by $\mathcal{S}_i$ occurred at time t_i, where n denotes the total number. For a narration $\mathcal{S}_i$ with timestamp t_i, we define the corresponding clip $\mathcal{C}_i$ with the following start t_i^s and end t_i^e timepoints:

$$\begin{cases} t_i^s = \text{Max}(t_i - \beta_i/2\alpha, t_{i-1}), \\ t_i^e = \text{Min}(t_i + \beta_i/2\alpha, t_{i+1}), \end{cases} \tag{1}$$

where β_i is a parameter representing the average temporal distance between consecutive narrations, calculated as $\sum_{j=0}^{n-1}(t_{j+1} - t_j)/n$. This value is calculated for each video individually, with α serving as a scale factor derived from the average of all β_i across the entire Ego-ExoClip dataset (α = 1.92 seconds). To ensure that a description does not span multiple clips, we utilize the preceding and following timestamps as boundary markers.

EgoIT

To enhance the instruction-following capabilities of MLLMs for downstream egocentric tasks, we introduce the EgoIT dataset, which comprises approximately 600K samples sourced from multiple distinct domains. All data samples are standardized into a uniform format. Each sample consists of four main components: the video path, task instruction, question, and answer. To ensure diversity in instructions, we integrate the "dataset description", "task description", and "instruction example" information into a prompt template, using GPT-4o[2] to generate ten distinct instructions per dataset.

Our instruction-tuning EgoIT dataset can be broadly categorized into three areas: 1) Action Recognition, which focuses on enhancing the identification of hand-object interactions. We utilize the EGTEA (Li, Liu, and Rehg 2021) and Something-Something-V2 (Goyal et al. 2017) datasets

[1] UID mapping associate narrations and corresponding videos.
[2] https://openai.com/index/hello-gpt-4o/.

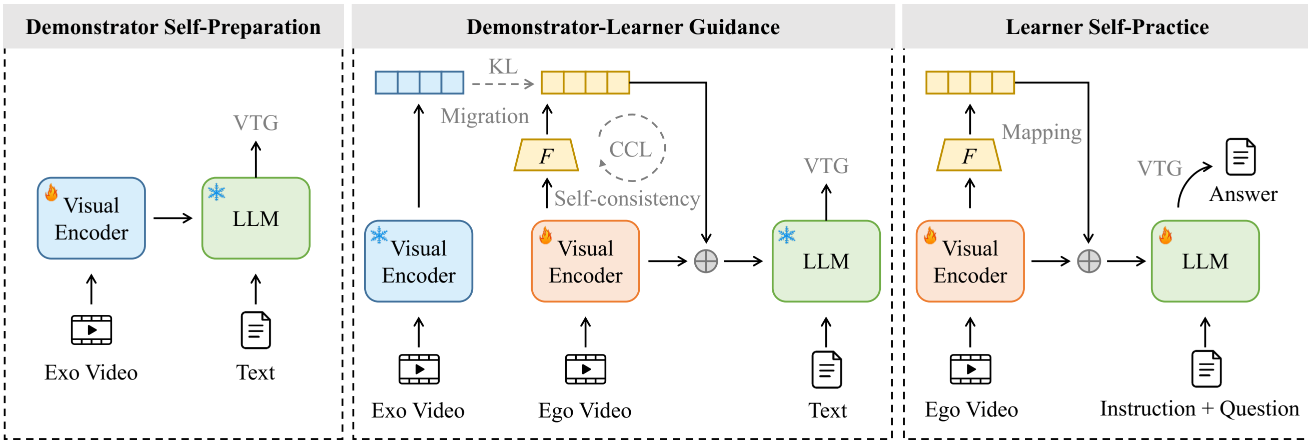

Figure 4: Illustration of our progressive training process, including the stages of Demonstrator Self-Preparation, Demonstrator-Learner Guidance, and Learner Self-Practice. “VTG”, “KL”, and “CCL” denote vision-grounded text generation, Kullback-Leibler divergence, and cycle consistency loss, respectively. F is a mapping function.

within the egocentric setting. 2) Question Answering, which aims to improve the ability to answer questions based on video content. We include the EgoTimeQA (Di and Xie 2024) and OpenEQA (Majumdar et al. 2024) datasets, encompassing both open-ended and close-ended questions. And 3) Captioning, which pays attention to enhancing basic visual description capabilities, utilizing the EgoExoLearn (Huang et al. 2024) dataset for this purpose.

Method

Under limited data conditions, bridging the gap between exocentric and egocentric domains is essential for improving the performance of MLLMs on various egocentric tasks. To tackle this challenge, we propose a progressive multimodal training paradigm, as illustrated in Figure 4. We describe each step in detail below.

Initialization

The goal of the initialization is to equip both the exocentric and egocentric visual encoders with foundational capabilities to process their respective data types. We tackle this by training the encoders separately using data from both domains. For the exocentric visual encoder, we utilize a large-scale, web-crawled dataset comprising image-text and video-text pairs from various publicly accessible databases, as summarized in Table 1. For the egocentric visual encoder, we employ the EgoClip (Lin et al. 2022) dataset, which includes a wide range of first-person daily activities. During this training phase, the parameters of LLM are frozen, and we optimize only the two visual encoders. The trained parameters of visual encoders then serve as initialization weights for the subsequent three-stage process.

Stage 1: Demonstrator Self-Preparation

After initialization, the exocentric visual encoder (the demonstrator) must adapt to the Ego-ExoClip dataset, despite already being proficient at understanding exocentric videos. To achieve this, we implement a self-preparation step to ensure that the exocentric encoder is adequately prepared to transfer its knowledge to the egocentric branch. As shown in Figure 4, we freeze the LLM parameters and fine-tune the exocentric visual encoder using exocentric clip-text data from our proposed Ego-ExoClip dataset. For supervision, we employ Vision-grounded Text Generation (VTG) loss to ensure data accuracy.

Stage 2: Demonstrator-Learner Guidance

In the second stage, we aim to establish a robust correlation between the exocentric and egocentric data, effectively mapping the learner to the demonstrator. This mapping is essential, as exocentric knowledge can significantly enhance the comprehension of egocentric videos, given that human behaviors are largely perspective-invariant and independent of camera viewpoint. For this stage, we continue to adopt Ego-ExoClip as input data. With the exocentric visual encoder sufficiently trained, we freeze it to maintain its integrity. We also freeze the LLM while fine-tuning the egocentric visual encoder and the mapping functions F and G. The mapping $F : X \to Y$ is inherently under-constrained, prompting us to couple it with an inverse mapping $G : Y \to X$, where X and Y denote the egocentric and exocentric domains, respectively. Both mappings are treated as bijections, and we enforce this structural assumption by training them simultaneously while applying a Cycle Consistency Loss (CCL), defined as:

$$\begin{aligned}\mathcal{L}_{\text{CCL}}(F, G) = \mathbb{E}_x \left[\|G(F(x)) - x\|_1 \right] \\ + \mathbb{E}_y \left[\|F(G(y)) - y\|_1 \right],\end{aligned} \tag{2}$$

where $x \in X$, $y \in Y$, and $\|\|_1$ denotes the L1 norm. The CCL encourages both forward cycle consistency, i.e., $x \to F(x) \to G(F(x)) \approx x$, and backward cycle consistency, i.e., $y \to G(y) \to F(G(y)) \approx y$. Additionally, we introduce the Kullback–Leibler (KL) divergence between the exocentric sample y and the estimated exocentric sample

Process	Dataset	Total	
		Exo	Ego
Initialization	Panda-70M, WebVid-10M, VIDAL-10M, InternVid-10M CC-3M, DCI-7.8K, EgoClip-3.8M	103M	3.8M
Stage 1	Ego-ExoClip-1.1M	1.1M	-
Stage 2	Ego-ExoClip-1.1M	1.1M	261K
Stage 3	OpenEQA-1.6K, EgoTimeQA-303K, Something-Something-V2-221K EGTEA-15K, EgoExoLearn-45K	-	586K

Table 1: Overview of the datasets utilized in the multiple training stages, detailing their sources and scales.

Method	ES	QAEgo4D		EgoTaskQA		CE	EK	EP	VLN-QA	EgoMCQ	
	Acc.	Acc.	Acc./Score	Acc.	Acc.	mAP	mAP	Acc.	Acc.	Acc.	Acc.
Specific Egocentric Methods											
EgoVLP	-	22.6	17.3/1.9	42.5	38.7	25.0	26.0	-	-	-	-
EgoVLPv2	-	25.8	19.7/2.1	46.3	42.3	30.7	29.1	15.3	10.4	72.3	30.6
MFAS	-	37.1	23.2/2.2	48.7	45.4	37.4	30.3	17.2	9.5	77.5	34.2
GroundVQA	-	50.2	29.0/2.6	-	-	42.5	-	22.7	17.4	80.3	37.1
LaViLa++	-	-	-	-	-	40.8	-	20.9	11.0	75.4	33.7
Zero-shot Closed-source MLLMs											
Gemini 1.5-Pro	71.2	-	-	-	-	-	-	-	-	-	-
GPT-4o	72.2	60.3	29.1/3.0	43.2	47.5	59.4	45.3	36.8	34.0	86.6	38.8
Zero-shot Open-source MLLMs											
LLaMA-VID	38.5	38.6	23.5/2.0	29.0	31.6	31.0	46.1	27.2	32.3	56.5	23.5
Video-LLaVA	38.4	35.2	**28.7**/2.7	29.8	30.6	43.5	45.2	27.5	34.0	60.9	25.1
VideoLLaMA2	51.7	48.5	25.3/2.6	42.1	42.6	68.1	47.8	29.3	37.5	82.1	32.0
VideoChat2	54.4	45.4	21.9/1.7	42.3	47.4	57.0	42.6	26.1	35.0	85.1	37.6
Qwen2.5-VL	57.6	55.3	26.4/2.8	42.6	45.8	64.2	46.4	32.5	34.3	85.3	38.2
Exo2Ego	**61.3**	**62.1**	28.3/2.7	**44.7**	**50.3**	**70.9**	**49.7**	**42.7**	**44.5**	**88.4**	**41.2**

Table 2: Performance comparison with recent state-of-the-art methods methods on eight benchmarks, where ES, CE, EK, and EP refer to EgoSchema, Charades-Ego, EPIC-KITCHENS-100, and EgoPlan. The QAEgo4D dataset is categorized into *close* and *open* question answering, EgoTaskQA includes *direct* and *indirect* subsets, and EgoMCQ contains *inter* and *intra* settings. The best results among open-source MLLMs are highlighted in bold.

$\hat{y} = F(x)$, aligning real and estimated exocentric feature distributions. The concatenated guidance from the demonstrator and the learner's own information is then fed into the LLM.

Stage 3: Learner Self-Practice

In the final stage, we focus on enhancing the instruction-following capabilities of MLLMs to address diverse egocentric downstream tasks. To achieve this, we employ the instruction-tuning data EgoIT. By concatenating the representations of the egocentric sample x and its corresponding mapped exocentric sample $F(x)$, we create a comprehensive semantic input for the LLM. To ensure the LLM aligns its responses more effectively with task instructions, we apply Low-Rank Adaptation (LoRA) to the frozen LLM, tuning it alongside the visual encoder and the mapping function F using VTG loss.

Experiment

Implementation Details

Datasets. The datasets employed in each stage of training are summarized in Table 1, showcasing a substantial amount of vision-text data across various types. After training, we evaluated our model on multiple benchmarks, which includes eight tasks with multiple configurations: EgoSchema (Mangalam, Akshulakov, and Malik 2023) for reasoning, QAEgo4D (Bärmann and Waibel 2022) for episodic memory, Charades-Ego (Sigurdsson et al. 2018) for action recognition, EPIC-KITCHENS-100 (Damen et al. 2022) for multi-instance retrieval, VLN-QA (Krantz et al. 2020) for egocentric navigation, EgoPlan (Chen et al. 2023) for planning, EgoMCQ (Lin et al. 2022) for alignment, and EgoTaskQA (Jia et al. 2022) for hybrid tasks.

Model. Our model is built upon the VideoLLaMA2 (Cheng et al. 2024) baseline. We consistently utilized the CLIP-Large-336 as the visual encoders and Mistral-7B-Instruct as the LLM. The mapping functions F and G consist of nine ResNet blocks between the down-sampling and up-sampling operations. In Stage 3, we incorporated LoRA into the LLM, with a rank set to 128, an alpha value of 256, and a dropout rate of 0.1. All experiments are conducted using the PyTorch framework on 16 A800 GPUs.

Baselines. We selected three categories of methods as comparative baselines, including specific approaches tailored for the egocentric domain (e.g., GroundVQA (Di and Xie 2024) and LaViLa++ (Xu et al. 2025)), closed-source MLLMs (e.g., GPT-4o), and open-source MLLMs (e.g., VideoL-LaMA2 (Cheng et al. 2024) and Qwen2.5-VL (Bai et al. 2025)). Except for LLaMA-VID (Li, Wang, and Jia 2023) and Video-LLaVA (Lin et al. 2023), which use fixed input settings of 1 fps and 8 frames respectively, all other models adopt a 16-frame configuration to ensure a fair comparison.

Quantitative Evaluation

Tables 2 presents the experimental results with different superior methods across various tasks and settings, which can be analyzed as follows:

- The experimental results on the reasoning and episodic memory tasks (i.e., the first three benchmarks) indicate that, in a zero-shot setting, MLLMs achieve performance comparable to specialized egocentric methods, highlighting their strong generalization capability across data domains. Notably, our Exo2Ego method significantly surpasses all open-source MLLMs and egocentric-specific methods under nearly all settings across the three egocentric evaluation benchmarks.
- The results on the fourth and fifth benchmarks present zero-shot performace on action recognition (Charades-Ego) and multi-instance retrieval (EPIC-KITCHENS-100). Our Exo2Ego method demonstrates consistent improvements over all baselines across all metrics for both tasks, validating its effectiveness in understanding the details of hand-object interactions.
- Planning and navigation are critical tasks in the domain of egocentric understanding. The sixth and seventh columns summarize our method's performance against existing baselines. The results reveal that open-source MLLMs exhibit accuracy rates concentrated between 26.0% and 32.5% on EgoPlan, highlighting their limitations for planning. In contrast, our method significantly outperforms all competitors, including GPT-4o (5.9% and 10.5% absolute gains), demonstrating its strong capability to effectively tackle diverse egocentric tasks.
- Video-text matching, a fundamental alignment task, is also crucial for comprehensive evaluation. As shown in the final column, our Exo2Ego approach consistently achieves the best performance across both settings, further demonstrating its superiority in generating accurate visual descriptions.

Ablation Study

In this section, we conducted a comprehensive analysis of how our model architecture, parameter update, training data, and training strategy impact overall performance.

Model Architecture. As summarized in Table 3, we investigated the impact of the mapping functions, LoRA, and loss functions on model performance. Comparing results 1 and 2, we observed that the introduction of LoRA significantly improves instruction-following capabilities, leading to enhanced performance across various tasks. This highlights LoRA's effectiveness in refining the model's ability to follow complex instructions. The third experiment utilizes only forward cycle consistency, resulting in a performance

Num	Mapping F	Mapping G	LoRA	Loss KL	Loss CCL	Avg
1	✓	✓	✓	✓	✓	**55.6**
2	✓	✓	✗	✓	✓	53.2 ↓
3	✓	✗	✓	✓	⇨	54.9 ↓
4	✓	✗	✓	✓	✗	54.4 ↓
5	✓	✓	✓	✗	✓	51.4 ↓
6	✰	✰	✓	✓	✓	54.7 ↓

Table 3: Performance comparison of different model architecture. The LLM is fixed at Mistral-7B-Instruct-v0.2, "✰" denotes replacing the existing modules with fully connected networks, and "⇨" represents removing backward cycle consistency, using only the forward direction. "Avg" is the average performance across all metrics in Figure 1.

Num	Stage 2 Exo ViEnc	Stage 2 Ego ViEnc	Stage 3 Ego ViEnc	Avg
1	❄	🔥	🔥	**55.6**
2	🔥	🔥	🔥	54.9 ↓
3	❄	🔥	❄	52.5 ↓

Table 4: Effects of different parameter update settings.

decline and highlighting the importance of the full bidirectional loss design. In the comparison between results 1 and 4, we found that the mapping G and CCL loss play a critical role in establishing an effective mapping between the egocentric and exocentric domains, thereby improving performance on downstream egocentric tasks. In the fifth experiment, we removed the KL loss, which incorporates guidance from exocentric data. The performance gap between this result and result 1 reinforces the importance of exocentric knowledge in enhancing egocentric video understanding. In the sixth experiment, we replaced the existing mapping functions with fully connected networks, which led to a slight performance decline. This suggests that more complex network architectures are capable of learning improved mapping transformations.

Parameter Update. The choice of visual encoders plays a crucial role in determining overall model performance, as detailed in Table 4. In the second experiment, we simultaneously fine-tuned both the exocentric and egocentric visual encoders during Stage 2, allowing the teacher and student to learn from each other. Comparing results 1 and 2, we observed that fine-tuning both encoders together may lead to the establishment of a neutral mapping between the domains, which ultimately results in poorer downstream performance. The gap between results 1 and 3 emphasizes the importance of maintaining more learnable parameters for visual adaptation during instruction fine-tuning.

Training Data. To assess the influence of training data on the model performance, we conducted experiments using VideoLLaMA2 (Cheng et al. 2024) as a baseline, as summarized in Table 5. Notably, VideoLLaMA2 adopts the same backbone architecture as our framework, including visual encoder and LLM. In the second row of results, we added the EgoClip dataset, which is aligned with the egocentric

EgoClip	Ego-ExoClip	EgoIT	Avg
✗	✗	✗	38.9
✓	✗	✗	45.2 ↑
✓	✓	✗	47.8 ↑
✓	✓	✓	49.7 ↑
✓	✓	✓	**55.6 (Ours)**

Table 5: Effect comparison across different datasets. The results in the first four rows are from VideoLLaMA2, using the same backbone as our method.

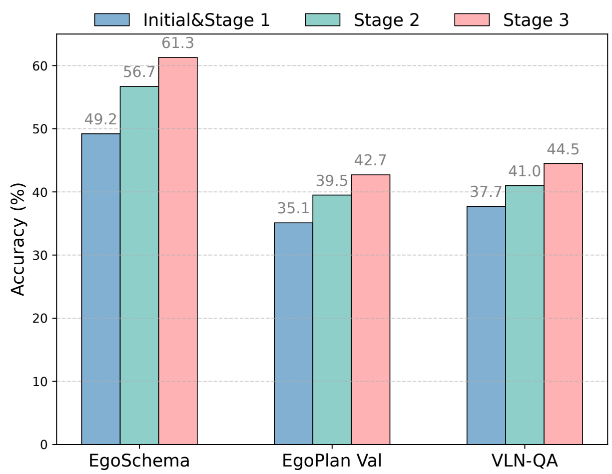

Figure 5: Accuracy comparison across different stages.

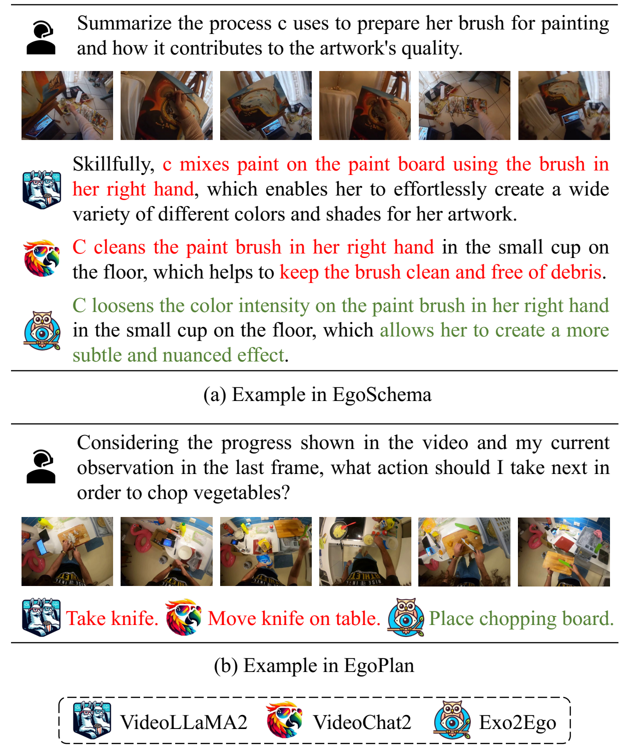

Figure 6: Two examples from different benchmarks.

domain. The improved performance suggests that training on data closely related to the downstream tasks leads to better domain knowledge. Compared to result 2, result 3 incorporates the Ego-ExoClip data and achieves further performance improvements. This validates the effectiveness of our constructed dataset. The fourth result indicates a significant improvement over the third result, demonstrating that our egocentric instruction-tuning data EgoIT enhances adaptation to downstream egocentric tasks. The performance gap between the last two rows indicates that, even when using the same datasets, our framework still exhibits a significant advantage. This demonstrates the superiority of our dual-encoder architecture and transfer-based training strategy.

Training Strategy. To quantify the contribution of each training stage, we evaluated each stage separately and the results are shown in Figure 5. The results show that pre-training establishes a foundation for basic egocentric video comprehension, while mapping learning in Stage 2 effectively aligns cross-domain data, yielding competitive performance. Moreover, the enhanced instruction-following capability achieved in Stage 3 further improves performance on downstream tasks. These results substantiate the efficacy of our three-stage training pipeline.

Qualitative Comparison

To visually compare the performance of different MLLMs, we presented several qualitative examples in Figure 6. In Figure 6(a), both baselines struggle with accurately interpreting the video content. VideoLLaMA2 erroneously assumes the camera-wearer is mixing paint with a brush, while VideoChat2 mistakenly believes the brush is being cleaned. In contrast, our Exo2Ego method correctly identifies that the camera-wearer is diluting the color on the brush, showcasing its superior understanding and inference capabilities for visual details. In the second example from EgoPlan dataset, which evaluates planning abilities, our model demonstrates a clear understanding of the temporal relationships between actions. Unlike the two baselines, which provide incorrect predictions, our model accurately identifies the correct sequence of actions required to achieve the goal.

Conclusion

In this paper, we propose Exo2Ego, a unified framework that significantly advances egocentric video understanding by cost-effectively transferring rich exocentric knowledge from MLLMs. To enable robust cross-view representation learning, we introduce Ego-ExoClip, a large-scale dataset comprising 1.1 million synchronized egocentric-exocentric clip-text pairs. Additionally, we curate EgoIT, an egocentric instruction-tuning dataset designed to further enhance instruction-following capabilities. By leveraging a novel exocentric-to-egocentric migration strategy and a progressive three-stage training pipeline, Exo2Ego effectively bridges the domain gap and promotes egocentric self-consistency. Extensive experiments on multiple benchmarks demonstrate that Exo2Ego consistently achieves substantial performance improvements.

Acknowledgments

This work is supported by Shenzhen Science and Technology Program, No.:KQTD20240729102207002; Jiangsu Science and Technology Major Program, No.:BG2024041; the National Natural Science Foundation of China, No.:62376140, No.:625B2065, and No.:U23A20315; the Science and Technology Innovation Program for Distinguished Young Scholars of Shandong Province Higher Education Institutions, No.:2023KJ128; the Special Fund for Taishan Scholar Project of Shandong Province; the Major Key Project of Pengcheng Laboratory; the Fundamental Research Funds for the Central Universities, No.:HIT.DZJJ.2025048.

References

Akiva, P.; Huang, J.; Liang, K. J.; Kovvuri, R.; Chen, X.; Feiszli, M.; Dana, K.; and Hassner, T. 2023. Self-supervised object detection from egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5225–5237.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

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*.

Bärmann, L.; and Waibel, A. 2022. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1560–1568.

Chen, Y.; Ge, Y.; Ge, Y.; Ding, M.; Li, B.; Wang, R.; Xu, R.; Shan, Y.; and Liu, X. 2023. EgoPlan-Bench: Benchmarking Egocentric Embodied Planning with Multimodal Large Language Models. *arXiv preprint arXiv:2312.06722*.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Chu, Q.; Zhang, H.; Feng, Y.; Liu, M.; Guan, W.; Wang, Y.; and Nie, L. 2025a. Technical report for ego4d long-term action anticipation challenge 2025. *arXiv preprint arXiv:2506.02550*.

Chu, Q.; Zhang, H.; Liu, M.; Feng, Y.; Shi, H.; and Nie, L. 2025b. Intention-Guided Cognitive Reasoning for Egocentric Long-Term Action Anticipation. *arXiv preprint arXiv:2508.01742*.

Dai, W.; Li, J.; LI, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 49250–49267. Curran Associates, Inc.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2020. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 4125–4141.

Damen, D.; Doughty, H.; Farinella, G. M.; Furnari, A.; Kazakos, E.; Ma, J.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 1–23.

Di, S.; and Xie, W. 2024. Grounded Question-Answering in Long Egocentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12934–12943.

Dou, Z.-Y.; Yang, X.; Nagarajan, T.; Wang, H.; Huang, J.; Peng, N.; Kitani, K.; and Chu, F.-J. 2024. Unlocking exocentric video-language data for egocentric video representation learning. *arXiv preprint arXiv:2408.03567*.

Feng, Y.; Zhang, H.; Liu, M.; Guan, W.; and Nie, L. 2025. Object-Shot Enhanced Grounding Network for Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24190–24200.

Feng, Y.; Zhang, H.; Xie, Y.; Li, Z.; Liu, M.; and Nie, L. 2024. ObjectNLQ@ Ego4D Episodic Memory Challenge 2024. *arXiv preprint arXiv:2406.15778*.

Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, 5842–5850.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; 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*, 18995–19012.

Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19383–19400.

Huang, Y.; Chen, G.; Xu, J.; Zhang, M.; Yang, L.; Pei, B.; Zhang, H.; Dong, L.; Wang, Y.; Wang, L.; et al. 2024. EgoExoLearn: A Dataset for Bridging Asynchronous Ego-and Exo-centric View of Procedural Activities in Real World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22072–22086.

Jia, B.; Lei, T.; Zhu, S.-C.; and Huang, S. 2022. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35: 3343–3360.

Jiang, N.; Liu, T.; Cao, Z.; Cui, J.; Zhang, Z.; Chen, Y.; Wang, H.; Zhu, Y.; and Huang, S. 2023. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9365–9376.

Johnson, M. 2015. Embodied understanding. *Frontiers in Psychology*, 6.

Krantz, J.; Wijmans, E.; Majumdar, A.; Batra, D.; and Lee, S. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 104–120. Springer.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, Y.; Liu, M.; and Rehg, J. M. 2021. In the eye of the beholder: Gaze and actions in first person video. *IEEE transactions on pattern analysis and machine intelligence*, 45(6): 6731–6747.

Li, Y.; Nagarajan, T.; Xiong, B.; and Grauman, K. 2021. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6943–6953.

Li, Y.; Wang, C.; and Jia, J. 2023. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Lin, K. Q.; Wang, J.; Soldan, M.; Wray, M.; Yan, R.; Xu, E. Z.; Gao, D.; Tu, R.-C.; Zhao, W.; Kong, W.; et al. 2022. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35: 7575–7586.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Luo, M.; Xue, Z.; Dimakis, A.; and Grauman, K. 2025. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In *European Conference on Computer Vision*, 407–425. Springer.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Majumdar, A.; Ajay, A.; Zhang, X.; Putta, P.; Yenamandra, S.; Henaff, M.; Silwal, S.; Mcvay, P.; Maksymets, O.; Arnaud, S.; Yadav, K.; Li, Q.; Newman, B.; Sharma, M.; Berges, V.; Zhang, S.; Agrawal, P.; Bisk, Y.; Batra, D.; Kalakrishnan, M.; Meier, F.; Paxton, C.; Sax, A.; and Rajeswaran, A. 2024. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16488–16498.

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*, 36: 46212–46244.

Millerdurai, C.; Akada, H.; Wang, J.; Luvizon, D.; Theobalt, C.; and Golyanik, V. 2024. Eventego3d: 3d human motion capture from egocentric event streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1186–1195.

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*, 5285–5297.

Ragusa, F.; Farinella, G. M.; and Furnari, A. 2023. Stillfast: An end-to-end approach for short-term object interaction anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3636–3645.

Ramakrishnan, S. K.; Al-Halah, Z.; and Grauman, K. 2023. Naq: Leveraging narrations as queries to supervise episodic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6694–6703.

Shen, Y.; and Elhamifar, E. 2024. Progress-aware online action segmentation for egocentric procedural task videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18186–18197.

Shi, Z.; Qiu, H.; Wang, L.; Meng, F.; Wu, Q.; and Li, H. 2024. Cognition Transferring and Decoupling for Text-supervised Egocentric Semantic Segmentation. *arXiv preprint arXiv:2410.01341*.

Shiota, T.; Takagi, M.; Kumagai, K.; Seshimo, H.; and Aono, Y. 2024. Egocentric action recognition by capturing hand-object contact and object state. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6541–6551.

Sigurdsson, G. A.; Gupta, A.; Schmid, C.; Farhadi, A.; and Alahari, K. 2018. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Truong, T.-D.; and Luu, K. 2024. Cross-view action recognition understanding from exocentric to egocentric perspective. *Neurocomputing*, 128731.

Wang, H.; Singh, M. K.; and Torresani, L. 2023. Ego-only: Egocentric action detection without exocentric transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5250–5261.

Wang, Q.; Zhao, L.; Yuan, L.; Liu, T.; and Peng, X. 2023. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3307–3317.

Wang, Y.; Liu, M.; Shao, R.; Zhang, H.; Wen, B.; Yang, F.; Gao, T.; Zhang, D.; and Nie, L. 2025. Time: Temporal-sensitive multi-dimensional instruction tuning and benchmarking for video-llms. *arXiv preprint arXiv:2503.09994*.

Wang, Y.; Liu, M.; Wei, Y.; Cheng, Z.; Wang, Y.; and Nie, L. 2022. Siamese alignment network for weakly supervised video moment retrieval. *IEEE Transactions on Multimedia*, 25: 3921–3933.

Wen, H.; Song, X.; Yin, J.; Wu, J.; Guan, W.; and Nie, L. 2024. Self-Training Boosted Multi-Factor Matching Network for Composed Image Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5): 3665–3678.

Xu, B.; Wang, Z.; Du, Y.; Song, Z.; Zheng, S.; and Jin, Q. 2025. Do Egocentric Video-Language Models Truly Understand Hand-Object Interactions? In *The Thirteenth International Conference on Learning Representations*.

Xue, Z. S.; and Grauman, K. 2023. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36: 53688–53710.

Zhang, H.; Liu, M.; Gao, Z.; Lei, X.; Wang, Y.; and Nie, L. 2021. Multimodal dialog system: Relational graph-based context-aware question understanding. In *Proceedings of the 29th ACM international conference on multimedia*, 695–703.

Zhang, H.; Liu, M.; Li, Y.; Yan, M.; Gao, Z.; Chang, X.; and Nie, L. 2023. Attribute-guided collaborative learning for partial person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 14144–14160.

Zhang, H.; Liu, M.; Li, Z.; Wen, H.; Guan, W.; Wang, Y.; and Nie, L. 2025. Spatial Understanding from Videos: Structured Prompts Meet Simulation Data. In *Advances in Neural Information Processing Systems*, 1–16.

Zhang, H.; Liu, M.; Liu, Z.; Song, X.; Wang, Y.; and Nie, L. 2024a. Multi-Factor Adaptive Vision Selection for Egocentric Video Question Answering. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, 59310–59328. PMLR.

Zhang, H.; Xie, Y.; Feng, Y.; Li, Z.; Liu, M.; and Nie, L. 2024b. HCQA@ Ego4D EgoSchema Challenge 2024. *arXiv preprint arXiv:2406.15771*.