

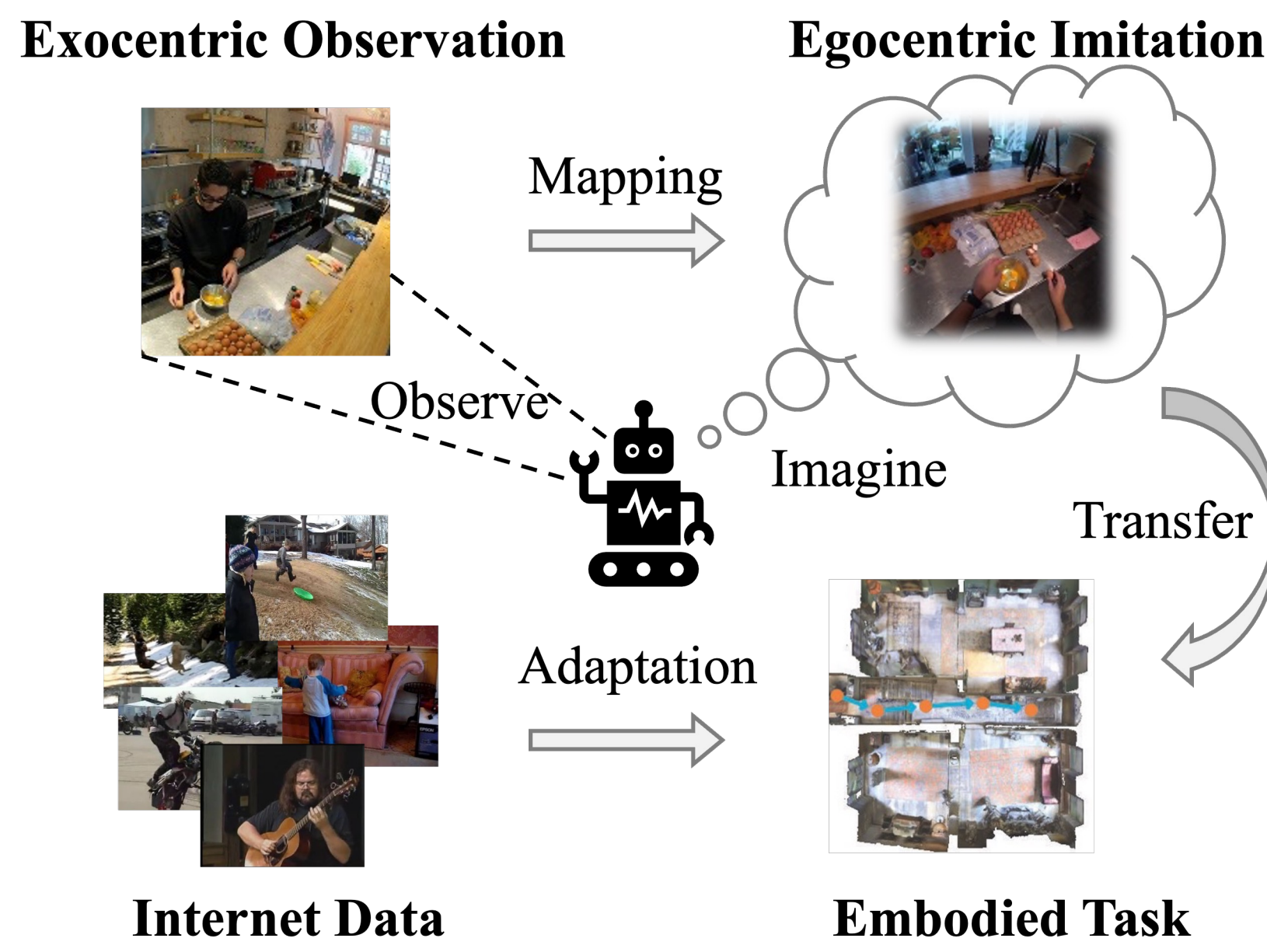
Exo2Ego: Exocentric Knowledge Guided MLLM for Egocentric Video Understanding

Haoyu Zhang^{1,2} Qiaohui Chu^{1,2} Meng Liu^{3,4} Haoxiang Shi^{1,2} Yaowei Wang^{1,2} Liqiang Nie¹

¹Harbin Institute of Technology (Shenzhen) ²Pengcheng Laboratory

³Shandong Jianzhu University ⁴Zhongguancun Academy

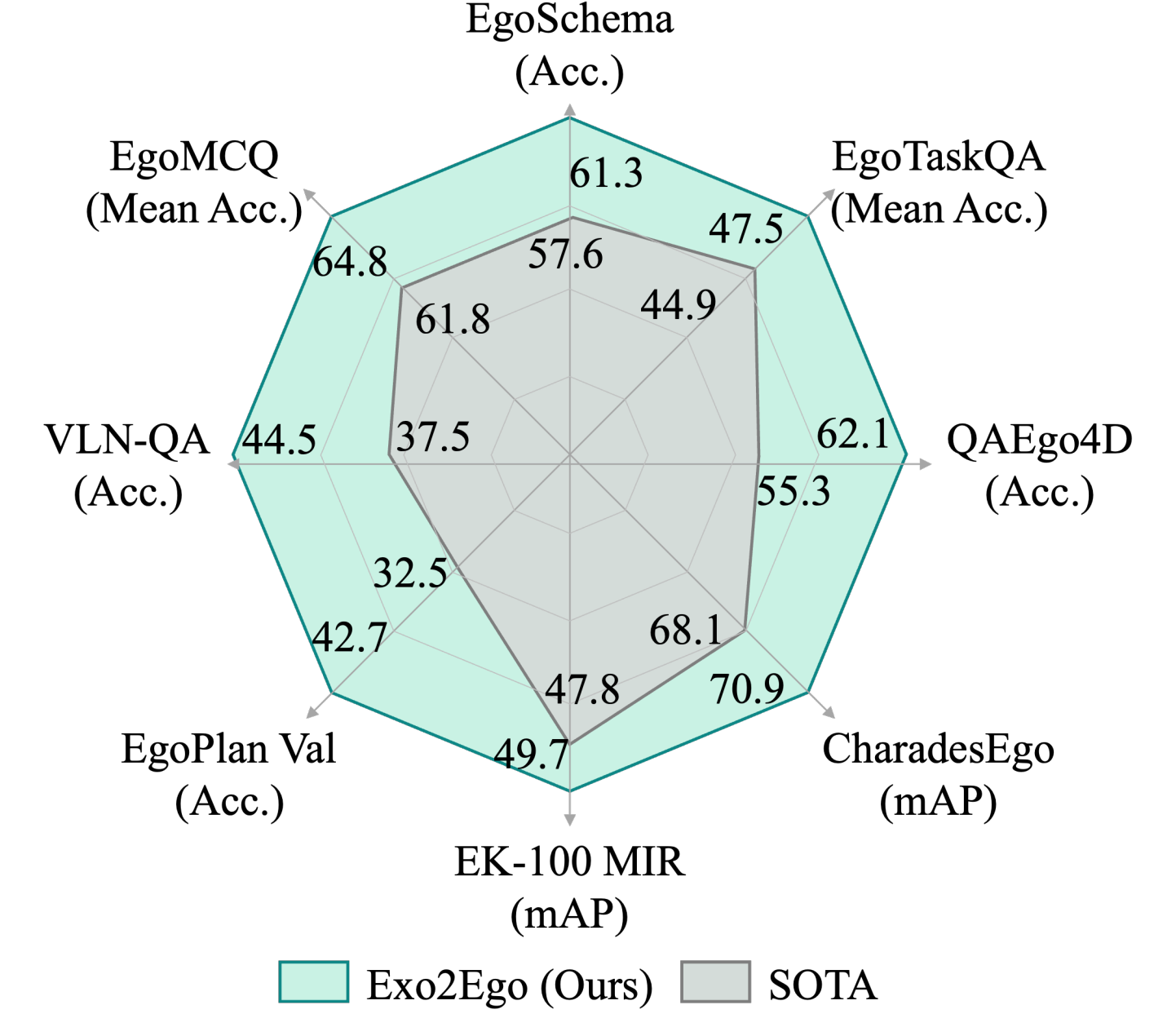
Motivation



- **Limited Egocentric Data:** Existing approaches are fundamentally constrained by the scarcity of egocentric video data, which hampers the effective training of large-capacity models.
- **Retrieval-Induced Misalignment:** Prior methods rely on retrieving exocentric videos to facilitate egocentric model training; however, this dependency introduces additional retrieval overhead and representation misalignment, ultimately leading to unstable performance.

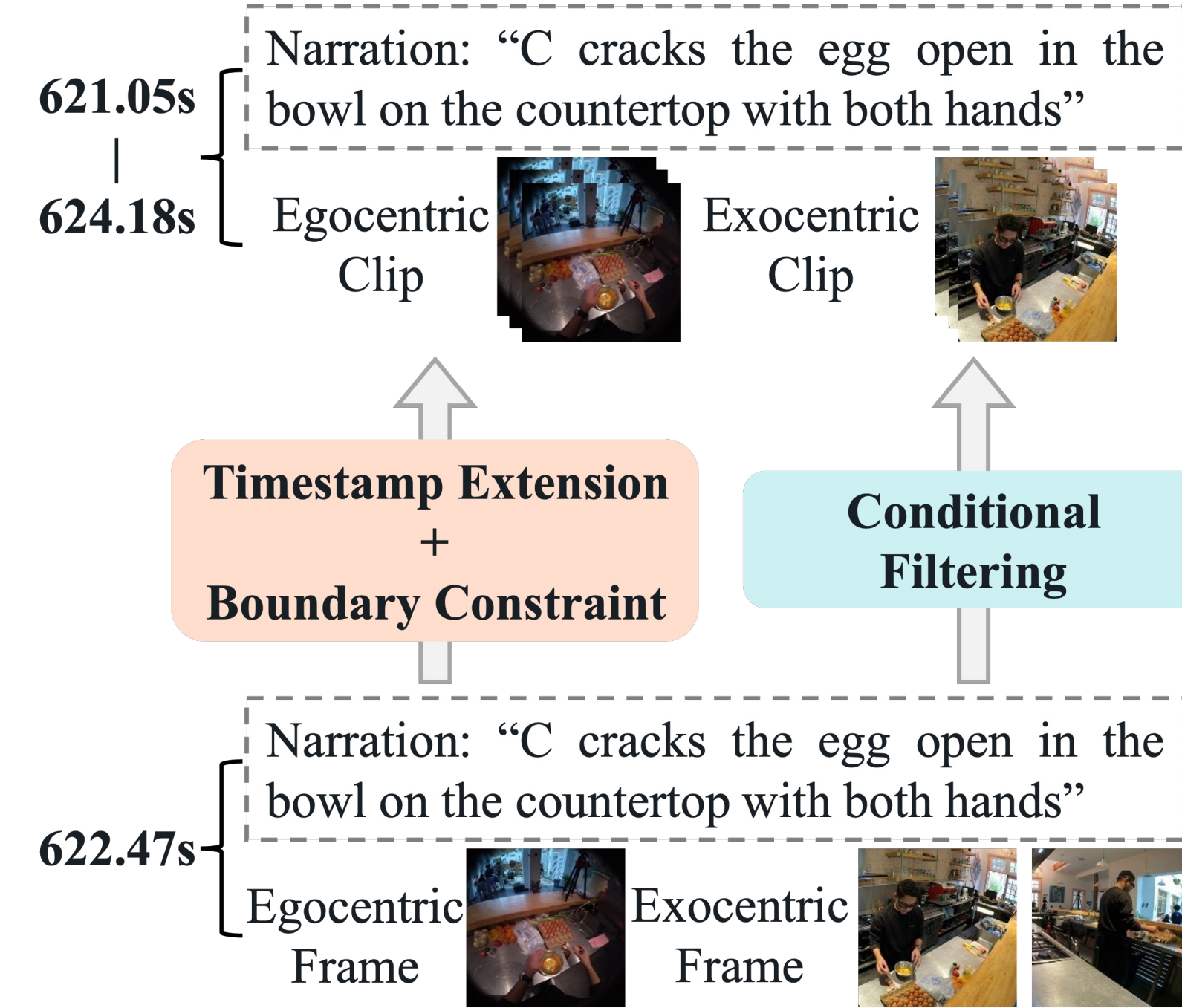
Drawing on findings from cognitive science showing that children learn by observing others' behavior (an exocentric view) and mapping these observations onto their own experiences (an egocentric perspective), we propose to establish a robust mapping between the two perspectives.

At-a-Glance



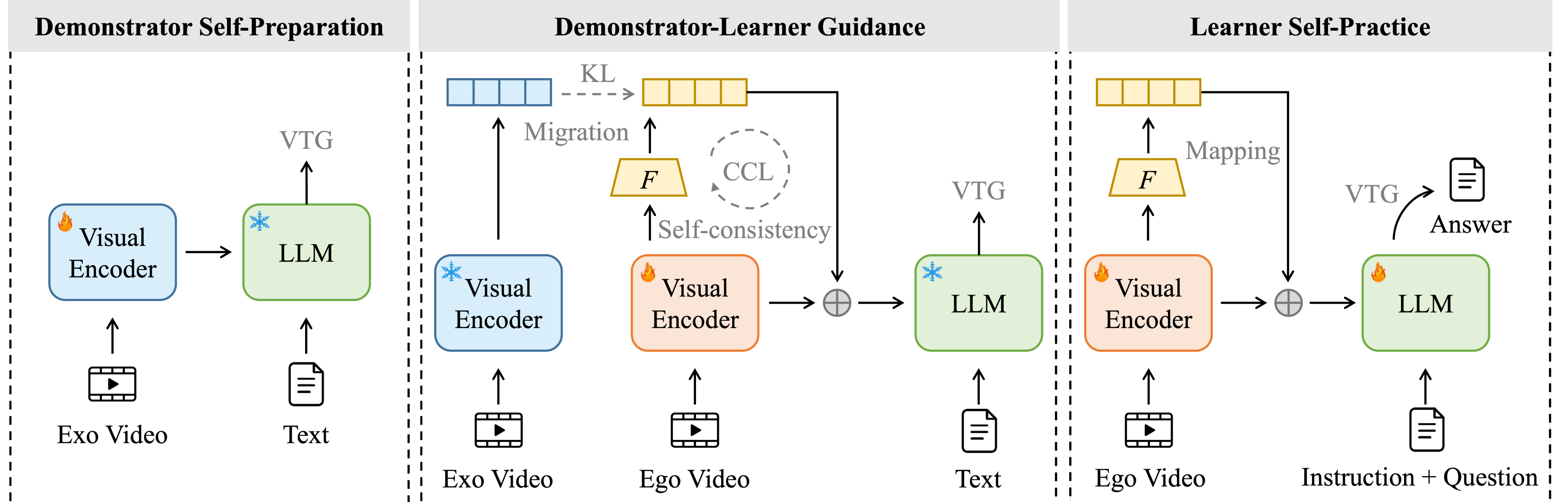
Dataset

We construct Ego-ExoClip, a pre-training dataset comprising 1.1M synchronized ego-exo pairs.



Framework

To bridge the gap between exocentric and egocentric domains, we propose a progressive multimodal training paradigm, which includes the stages of Demonstrator Self-Preparation, Demonstrator-Learner Guidance, and Learner Self-Practice.



Performance

Our method achieves superior performance across eight different benchmarks compared to a wide range of competing methods.

Method	ES Acc.	QA Ego4D Acc.	EgoTaskQA Acc./Score	CE mAP	EK mAP	EP Acc.	VLN-QA Acc.	EgoMCQ Acc.	Acc.
<i>Specific Egocentric Methods</i>									
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
MFAS	-	37.1	23.2/2.2	48.7	45.4	37.4	30.3	17.2	9.5
GroundVQA	-	50.2	29.0/2.6	-	-	42.5	-	22.7	17.4
LaViLa++	-	-	-	-	-	40.8	-	20.9	11.0
<i>Zero-shot Closed-source MLLMs</i>									
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
<i>Zero-shot Open-source MLLMs</i>									
LLaMA-VID	38.5	38.6	23.5/2.0	29.0	31.6	31.0	46.1	27.2	32.3
Video-LLaVA	38.4	35.2	28.7/2.7	29.8	30.6	43.5	45.2	27.5	34.0
VideoLLaMA2	51.7	48.5	25.3/2.6	42.1	42.6	68.1	47.8	29.3	37.5
VideoChat2	54.4	45.4	21.9/1.7	42.3	47.4	57.0	42.6	26.1	35.0
Qwen2.5-VL	57.6	55.3	26.4/2.8	42.6	45.8	64.2	46.4	32.5	34.3
Exo2Ego	61.3	62.1	28.3/2.7	44.7	50.3	70.9	49.7	42.7	44.5

Qualitative Comparison

To visually compare the performance of different MLLMs, we present two qualitative examples from different tasks.

Summarize the process c uses to prepare her brush for painting and how it contributes to the artwork's quality.

Skillfully, **c mixes paint on the paint board using the brush in her right hand**, which enables her to effortlessly create a wide variety of different colors and shades for her artwork.

C cleans the paint brush in her right hand in the small cup on the floor, which helps to **keep the brush clean and free of debris**.

C loosens the color intensity on the paint brush in her right hand in the small cup on the floor, which allows her to create a more subtle and nuanced effect.

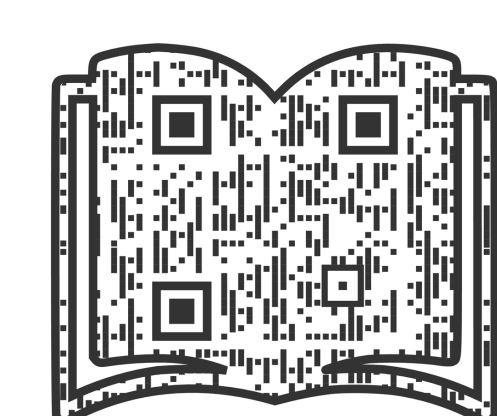
(a) Example in EgoSchema

Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to chop vegetables?

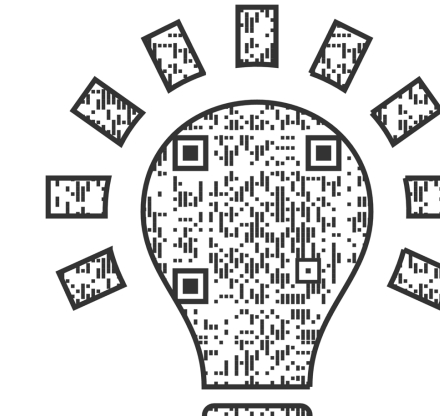
Take knife. Move knife on table. Place chopping board.

(b) Example in EgoPlan

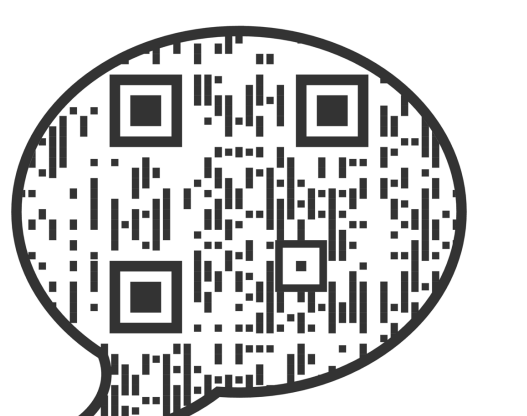
VideoLLaMA2 VideoChat2 Exo2Ego



Paper



Code



E-mail

Ablation Study

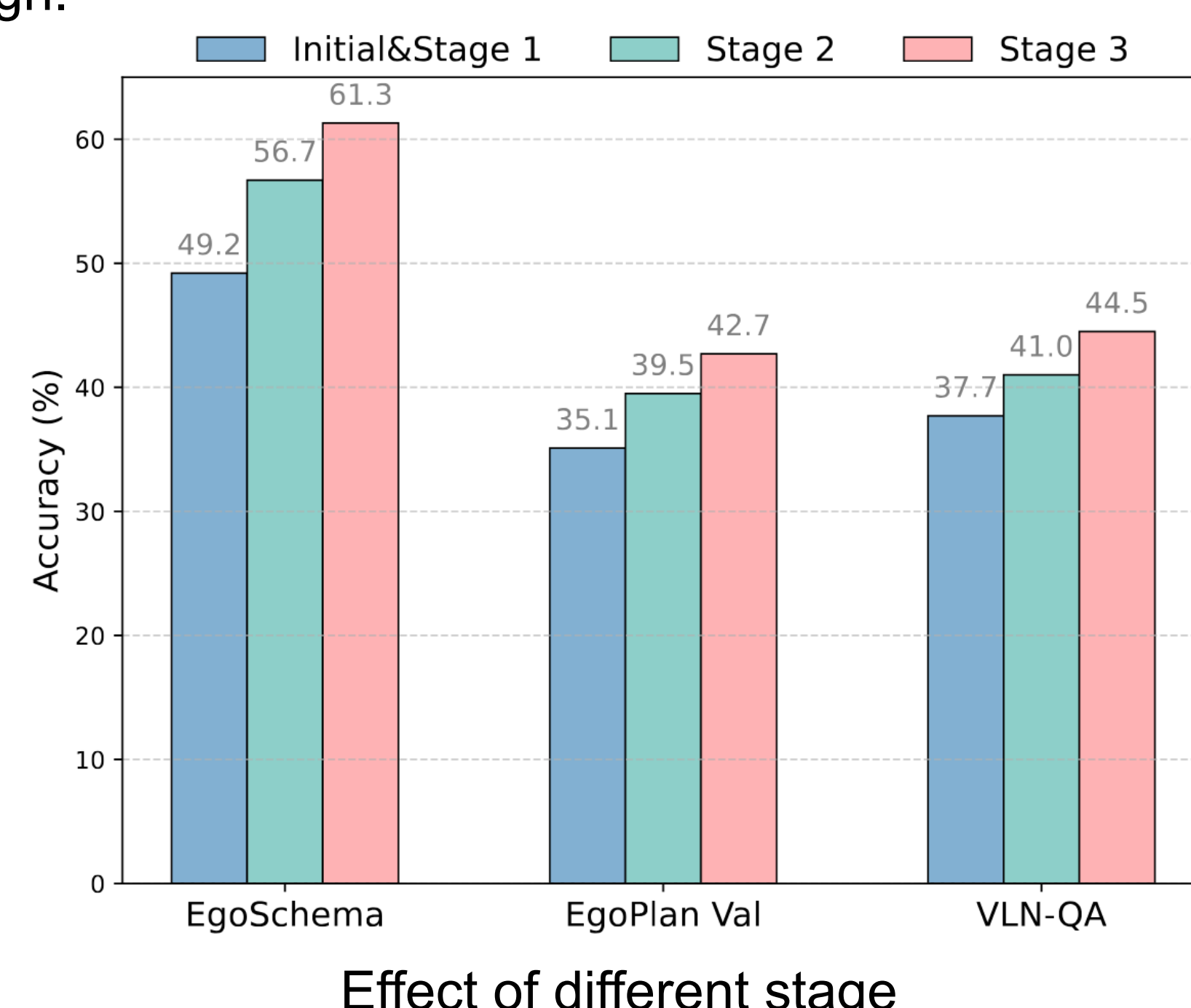
Through extensive ablation studies, we validate the effectiveness and soundness of each individual component as well as the overall architectural design.

Effect of different component

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 ↓

Effect of different data

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



Effect of different stage