

Learning to Perceive Egocentric Hand-Object Interactions

2025.4.27 by Boshen Xu at [NUS OMG Workshop](#)

AIM³ Lab, School of Information, Renmin University of China



AI · M³
中国人民大学多媒体计算实验室



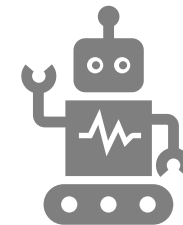
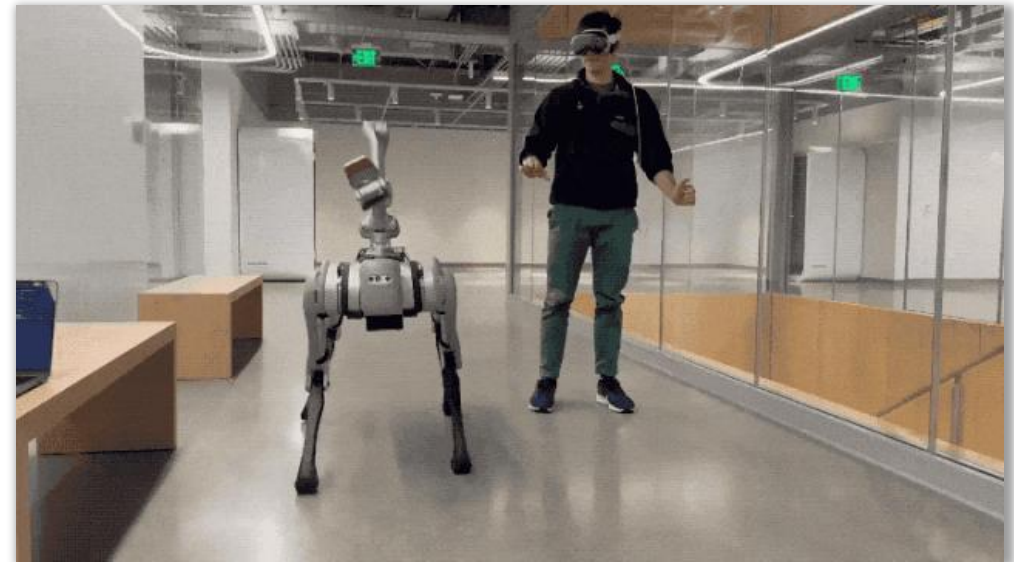
中國人民大學
RENMIN UNIVERSITY OF CHINA

🕶️ Egocentric AI: Perception and Interaction

Today's primary focus



Ego-Home



Ego-Robot

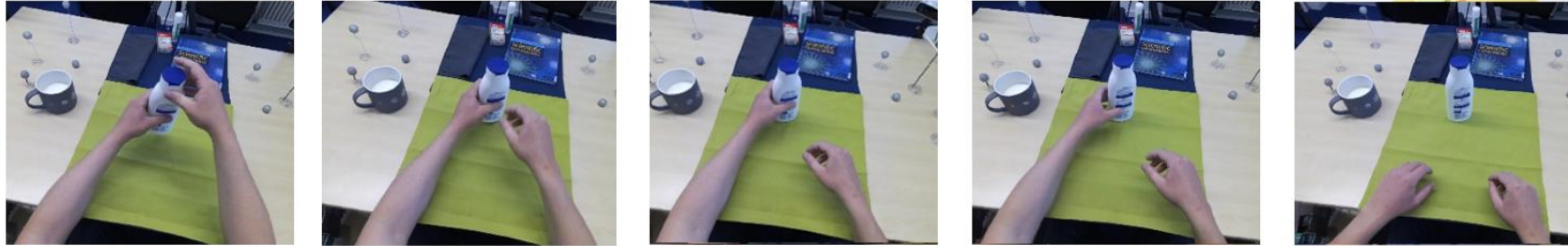
Humans Interact with Objects Everyday



Img source: [4]

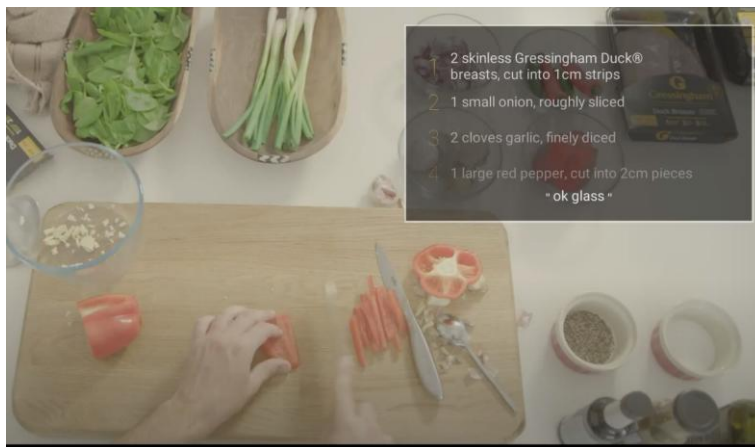
EgoHOI: Egocentric Hand-Object Interactions

Ego-HOI recognize how human interact with objects from a first-person view

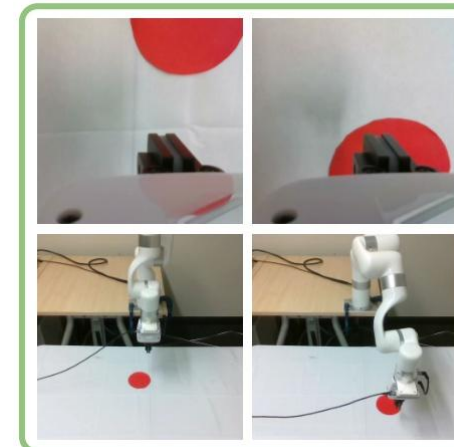


place lotion

VR / AR



Robot Manipulation



My Research: Learning to Perceive EgoHOI

Today's primary focus



POV [MM'23]

Learning EgoHOI from
Multi-View Observation

☹️ Simple Multi-choice Test

HOI-Verb ☹️

Video-Text Similarity Rank

1. C **hangs** the frying pan
2. C **whispers** the frying pan
3. C **writes** the frying pan
4. C **shakes** the frying pan
5. C **paints** the frying pan

GT: C shakes the frying pan

HOI-Noun ☹️

Video-Text Similarity Rank

1. C drops the **paper**
2. C drops the **cork**
3. C drops the **dough**
4. C drops the **spice**
5. C drops the **pan**

GT: C drops the dough

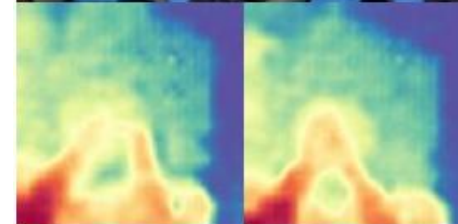
EgoNCE++ [ICLR'25]

Learning to Recognize Fine-
Grained EgoHOI from Text

Video



Ours



DAv2



EgoDTM [arxiv'25]

Learning EgoHOI from
Spatially-informed Supervision

Do Egocentric Video-Language Models Truly Understand Hand-Object Interactions?

Boshen Xu, Ziheng Wang, Yang Du, Zhinan Song, Sipeng Zheng, Qin Jin*

AIM³ Lab, School of Information, Renmin University of China



AI · M³
中国人民大学多媒体计算实验室



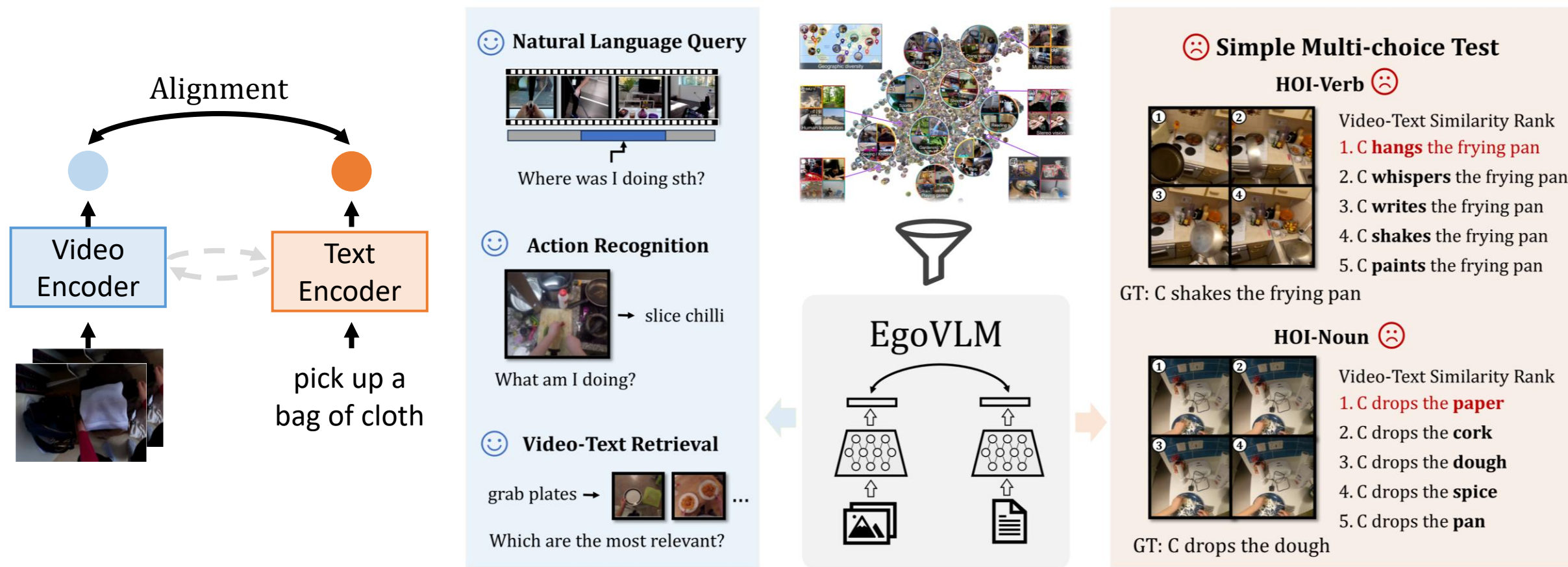
中國人民大學
RENMIN UNIVERSITY OF CHINA



ICLR

Observation: EgoVLMs Fails Simple Multi-Choice Test

- EgoVLMs are trained via prevailing VLP paradigm, but fails simple multi-choice test



Question: Do EgoVLM Truly Understand EgoHOIs?

- EgoHOIBench: 29K multi-choice test for EgoHOI recognition from Ego4D

Task form: multi-choice by video-text similarity



video

GT: *C opens an oven*

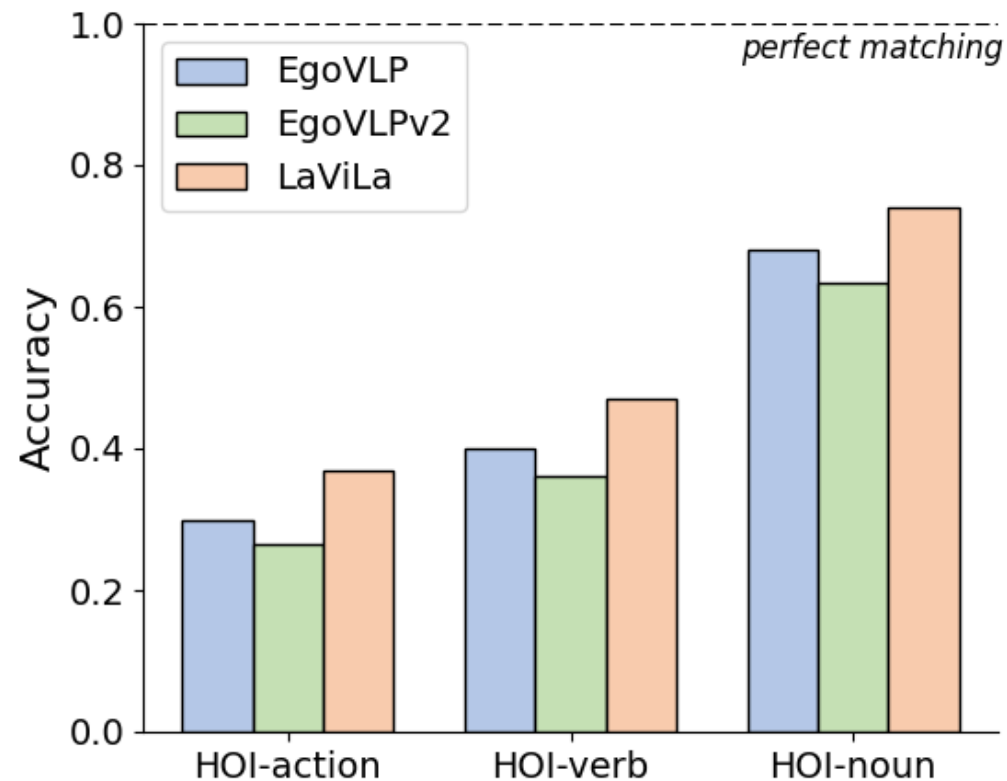
HOI-verb: *C {10 random verb candidates} an oven*
e.g., *C close an oven, C drops an oven,...*

HOI-noun: *C opens an {10 random noun candidates}*
e.g., *C opens a book, C opens a door,...*

HOI-action = correct HOI-verb + correct HOI-noun

Findings:

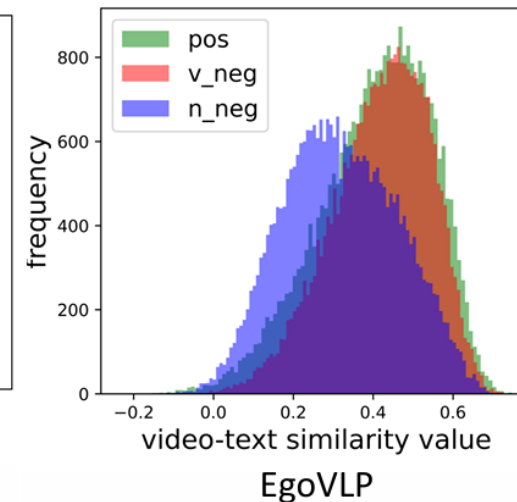
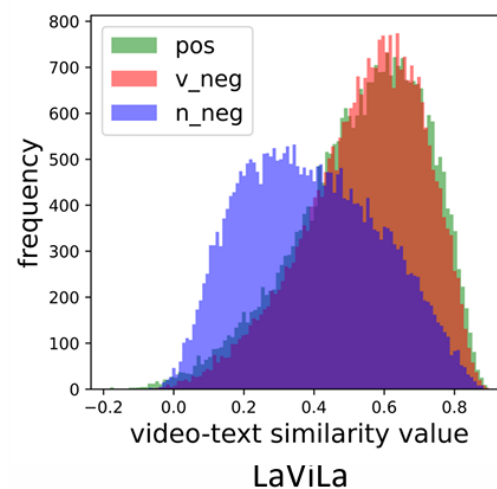
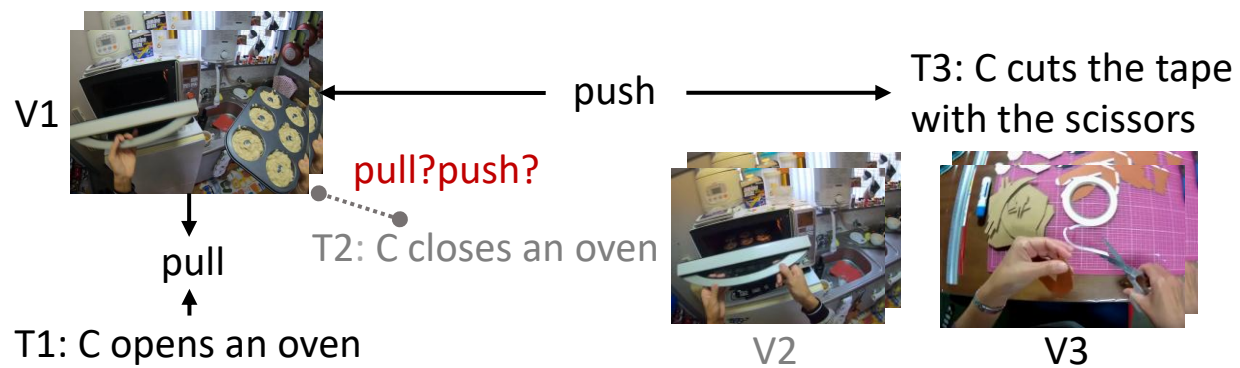
- EgoVLMs fail the simple multi-choice test. Why is that?
- HOI-noun recognition is better than that on HOI-verb. Why is that?



Analyses of Performance on EgoHOIBench

1. Why EgoVLMs fail the simple multi-choice test?
2. Why performance on HOI-noun is better than that on HOI-verb?

- Lack of fine-grained V2T supervision

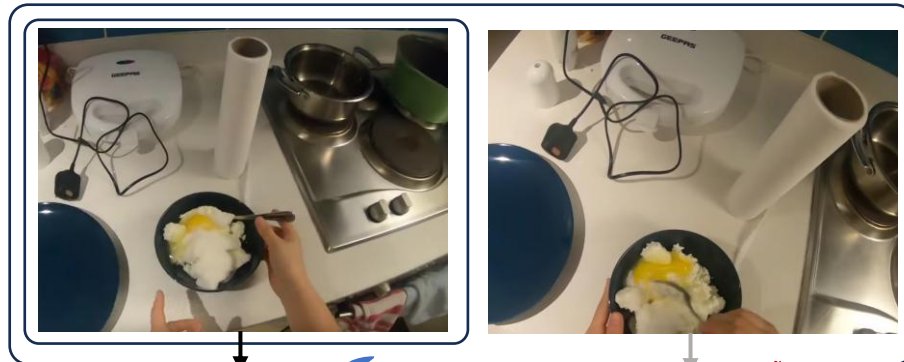


→ Can't distinguish T1 (pos) & T2 (v_neg) for V1!

→ HOI-noun is easier than HOI-verb

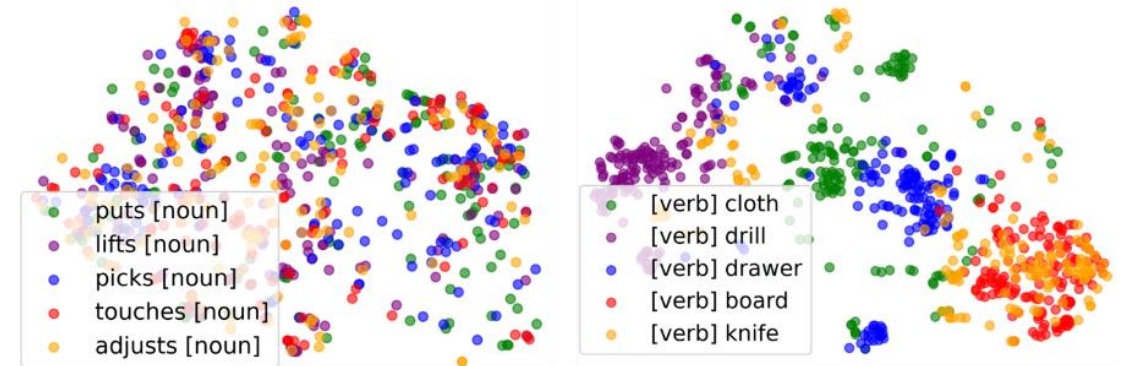
Analyses of Performance on EgoHOIBench

1. Why EgoVLMs fail the simple multi-choice test?
 2. Why performance on HOI-noun is better than that on HOI-verb?
- Bias towards understanding (static)nouns instead of (temporal)verbs



pull ✓
C stirs an egg

pull ✗
C C stirs an egg (subtle difference)



verb-anchored

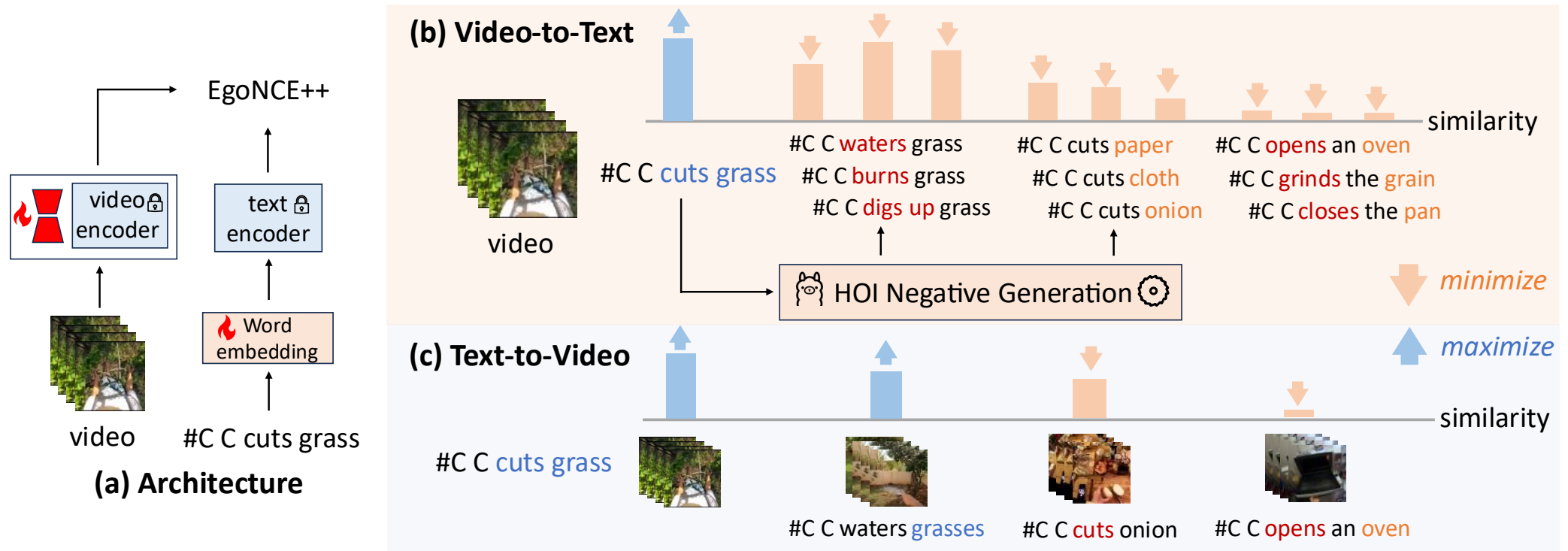
noun-anchored

t-SNE visualization

- EgoVLM learns an **object-centric feature space**
- HOI-noun recognition is more **linear separable** than that on HOI-verb

Can we solve the test on EgoHOIBench while leveraging the object-centric nature of feature space via a unified objective?

EgoNCE++: Asymmetric Contrastive Learning Objective



- Video-to-Text: enrich the hard negative supervision

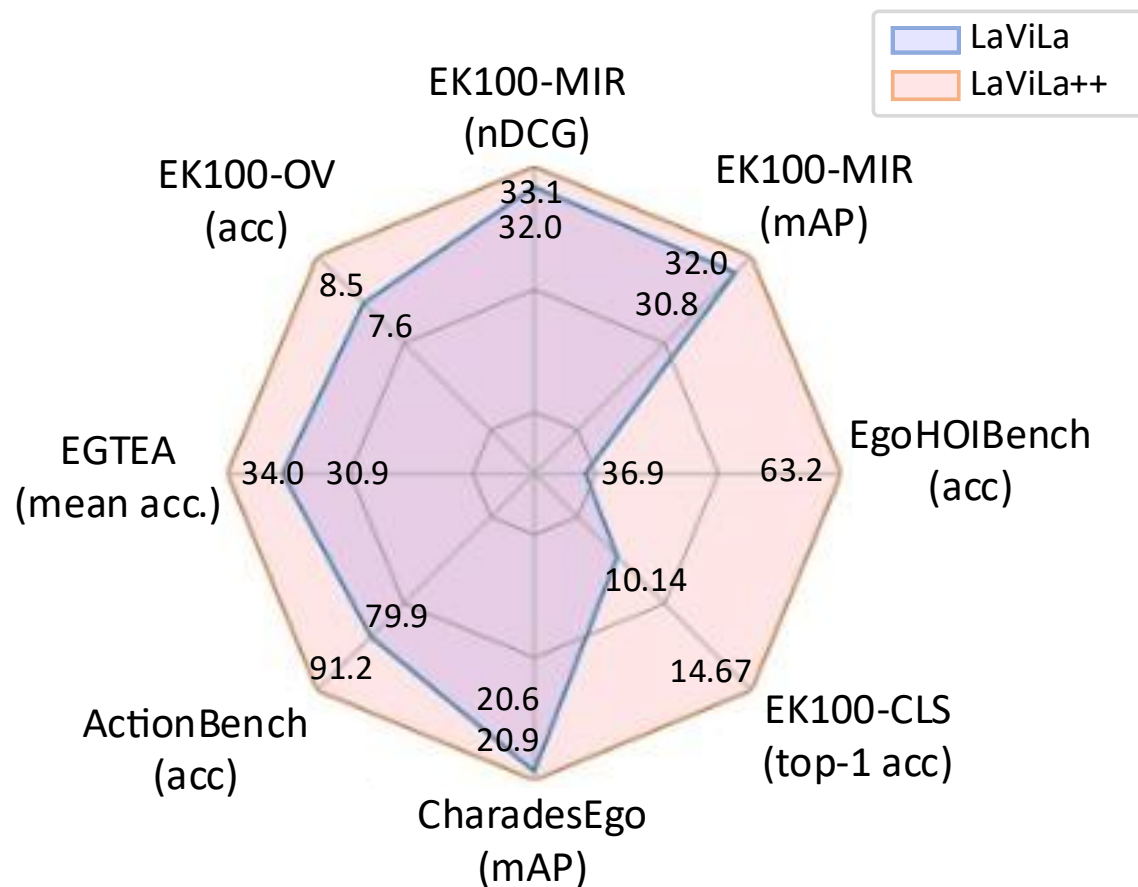
$$\mathcal{L}_{v2t} = \frac{1}{B} \sum_{v_i \in \mathcal{B}(v)} \log \frac{\exp(v_i \cdot t_i / \tau)}{\sum_{t_j \in \mathcal{B}(t)} \exp(v_i \cdot t_j / \tau) + \sum_{t_k \in \mathcal{N}_{\text{noun}}(t_i) \cup \mathcal{N}_{\text{verb}}(t_i)} \exp(v_i \cdot t_k / \tau)}$$

- Text-to-Video: maintain the object-centric feature space

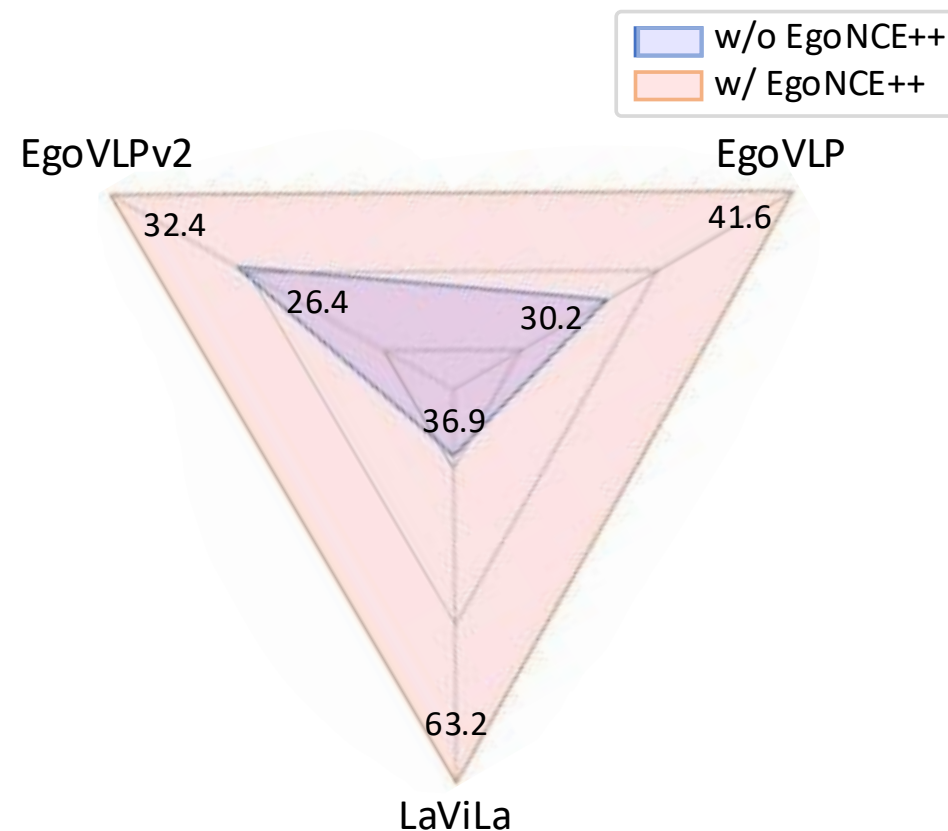
$$\mathcal{L}_{t2v} = \frac{1}{B} \sum_{t_i \in \mathcal{B}(t)} \log \frac{\sum_{k \in \mathcal{P}_n(v_i)} \exp(t_i \cdot v_k / \tau)}{\sum_{v_j \in \mathcal{B}(v)} \exp(t_i \cdot v_j / \tau)}$$

Consistently Better Generalization in Zero-Shot Settings

- Performance on EK-100-MIR: comparable to HelpingHands, HENASSY

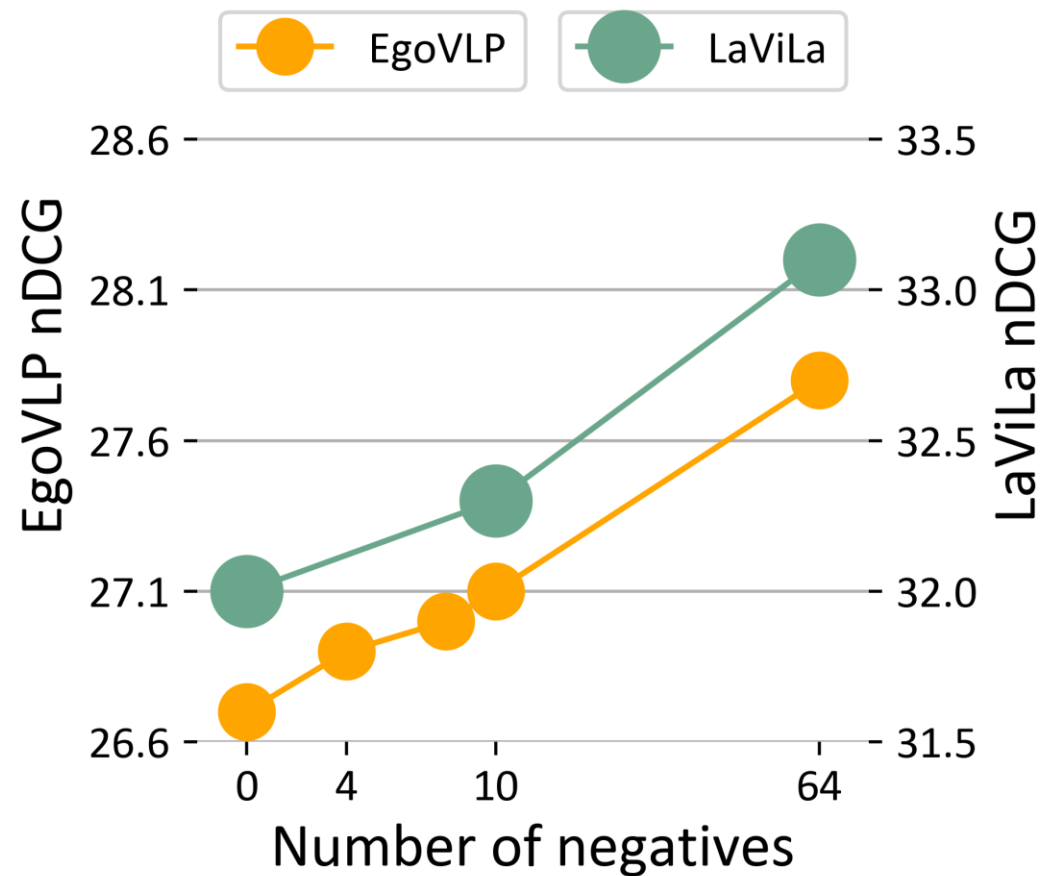
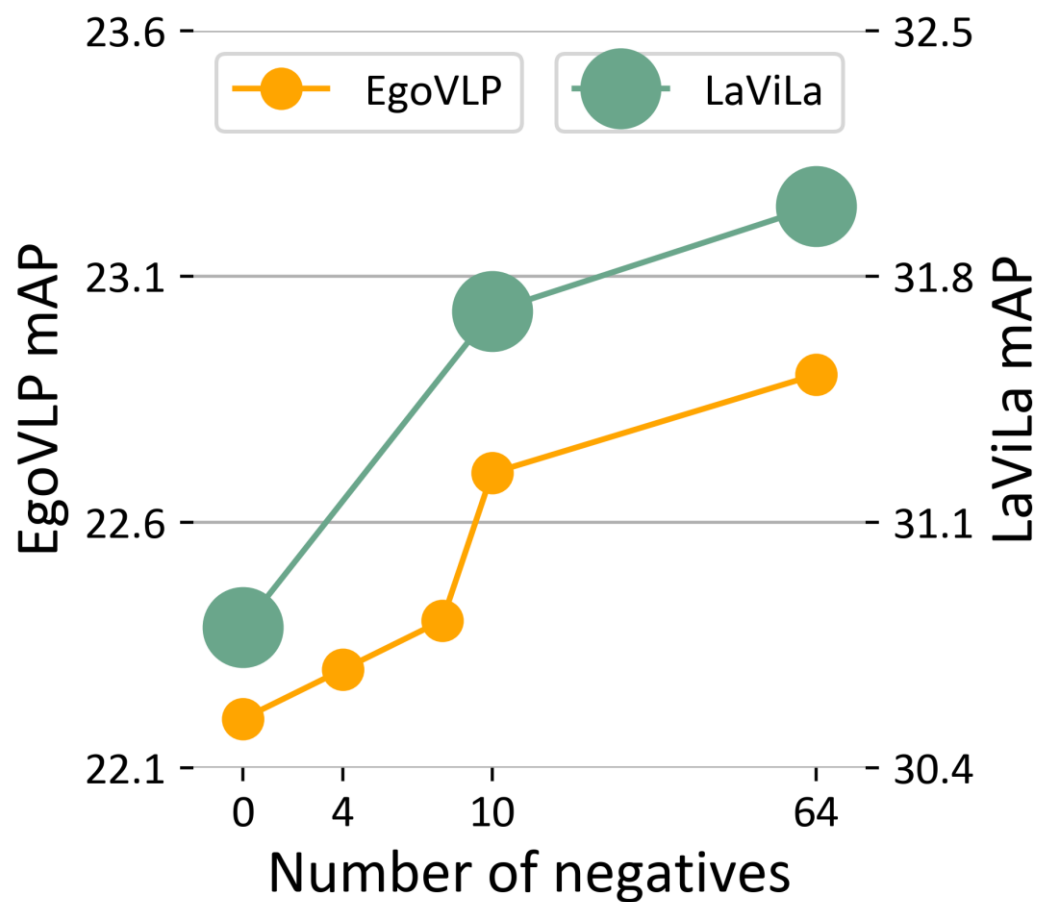


(a) Across Benchmarks



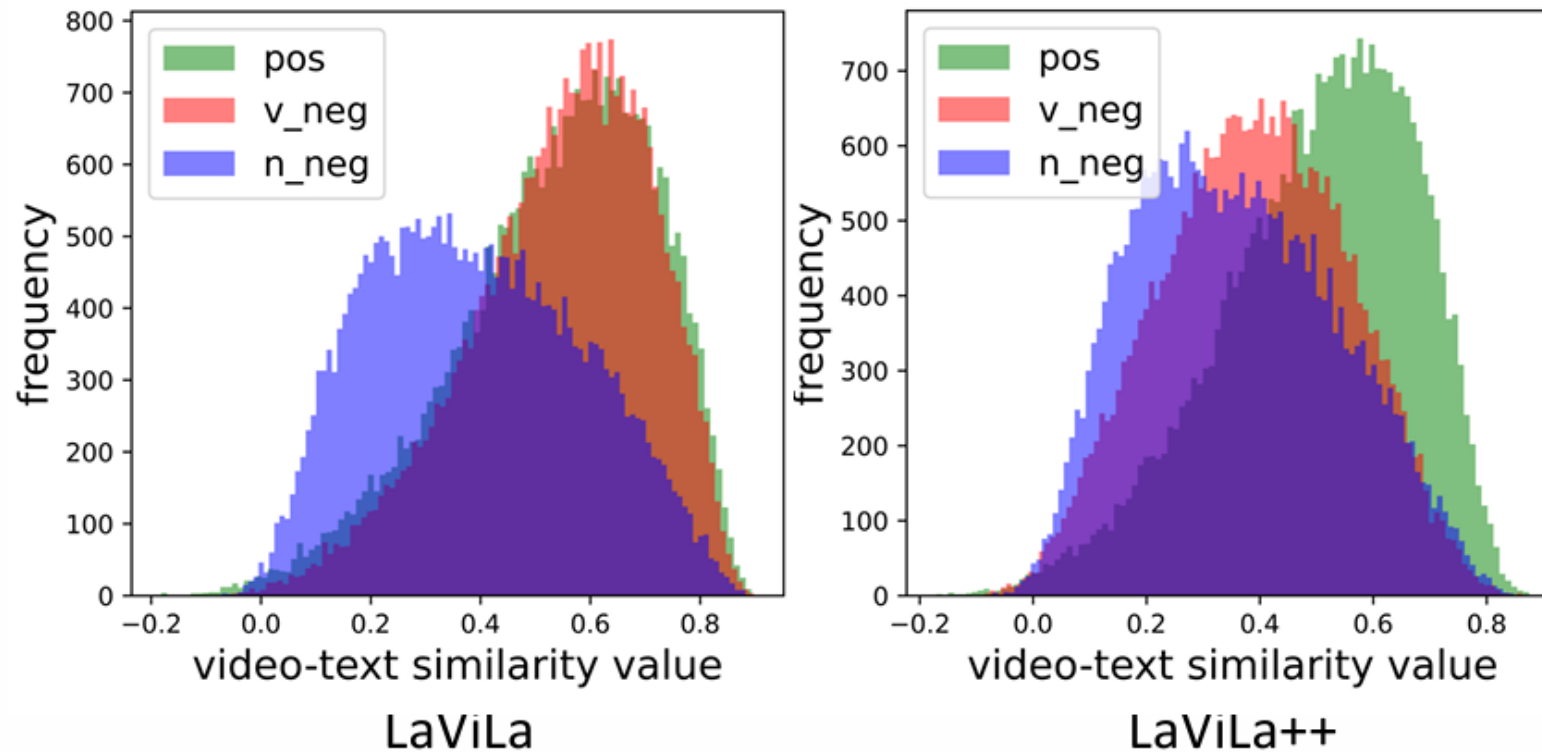
(b) Across EgoVLMs

Ablation of Scaling Negative Texts



EgoNCE++ Improves Video-Text Similarity Distribution

- Histogram of video-text similarity
 - Sim of (Video, Negative Verbs) decreases
 - Sim of (Video, Positives) retains



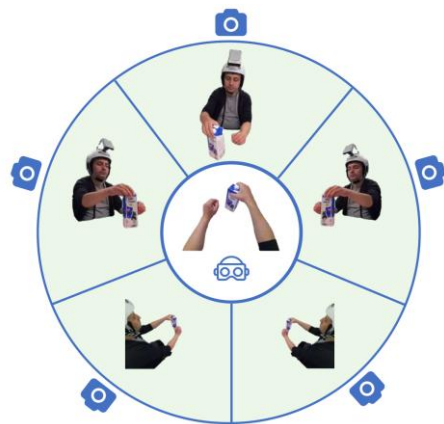
Takeaways

Paper Contributions

- **EgoHOIBench**: a simple multi-choice testbed for EgoHOI understanding
- **EgoNCE++**: an EgoHOI-aware asymmetric video-language pretraining objective
- More **generalizable EgoVLMs** across **downstream benchmarks**

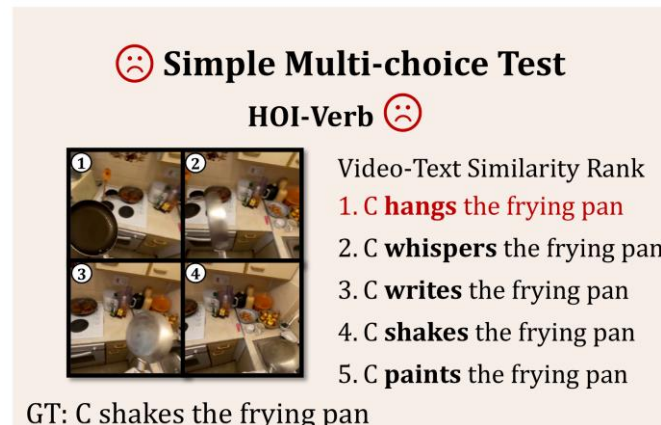
My Previous, Current, and Future Research

Previous work: EgoHOI short video understanding / VLP



POV [MM'23]

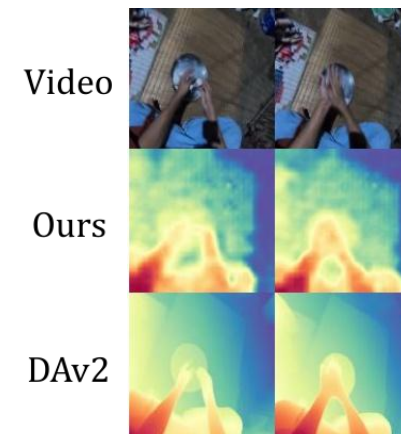
3rd to 1st view transfer



GT: C shakes the frying pan

EgoNCE++ [ICLR'25]:

fix EgoVLM on EgoHOI recognition



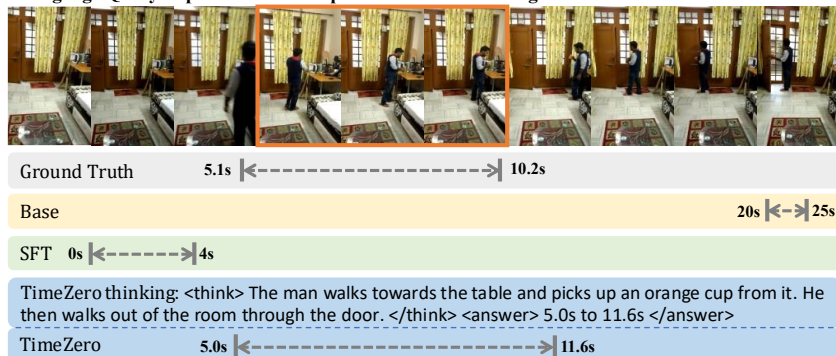
EgoDTM [arxiv'25]

3D-aware EgoVLM

↓ more temporal info

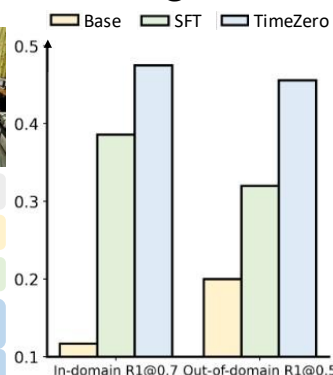
Current work: Long video understanding

Language Query: a person takes a cup from a table video length: 23s



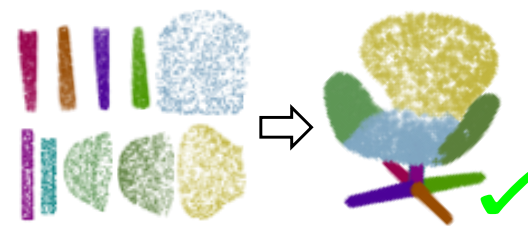
TimeZero [CVPRW'25]

a reasoning-guided MLLM for video grounding



↓ more spatial info

Future work: RoboHOI manipulation



SPAFormer [3DV'25]

object assembly task



Img source: [6]

Thank You!

If any questions, feel free to contact

boshenx@ruc.edu.cn,

or visit

<https://xuboshen.github.io/>

to check details in my papers!



AI · M³

中国人民大学多媒体计算实验室



中國人民大學

RENMIN UNIVERSITY OF CHINA



ICLR

References

- [1] page2: Demo of Apple Vision Pro, <https://www.apple.com/sg/apple-vision-pro/>
- [2] page2: VisionProTeleop, <https://github.com/Improbable-AI/VisionProTeleop>
- [3] page2: Ego-Home image is from Egocentric Survey [IJCV'24]
- [4] page3: Videos are from one slides by Yufei Ye, <https://judyye.github.io/ihoi/>
- [5] page4: robot manipulation image is from R3M [CoRL'22], EgoHOI image from H2O [ICCV'21]
- [6] page16: robot manipulation image is from ViViDex [ICRA'25]