From My View to Yours: Learning Egocentric Cues from Exocentric Video using Privileged Egocentric Supervision

Dominick Reilly[1] **Manish Kumar Govind**[1] **Le Xue**[2] **Srijan Das**[1]

[1] University of North Carolina at Charlotte, Charlotte NC, USA
[2] Elorian AI, Palo Alto CA, USA

https://github.com/dominickrei/EgoExo4ADL

dreilly1@charlotte.edu

Abstract. Vision Language Models (VLMs) have achieved strong performance across diverse video understanding tasks. However, their viewpoint invariant training limits their ability to understand egocentric properties (e.g., human-object interactions) from exocentric video observations. This limitation is critical for applications such as Activities of Daily Living (ADL) monitoring, where understanding egocentric properties is essential yet egocentric cameras are impractical to deploy, making it impossible to simply collect egocentric data at test time. To address this challenge, we propose Ego2ExoVLM, a VLM framework that learns to infer egocentric properties from exocentric videos. Our key insight is that time-synchronized ego-exo video pairs can be leveraged during training, where the egocentric viewpoint provides *privileged* supervision (rich egocentric signal available only at training time). Ego2ExoVLM achieves this through two components: **Ego2Exo Sequence Distillation**, which transfers egocentric reasoning through a language-level sequence distillation objective, and **Ego Adaptive Visual Tokens**, which encourages the model to surface relevant interaction cues within exocentric video representations. To measure this capability, we introduce **Ego-in-Exo Perception**, a benchmark designed to evaluate the understanding of egocentric properties from exocentric videos. Ego2ExoVLM is evaluated on 10 tasks across Ego-in-Exo Perception and existing ADL benchmarks, achieving state-of-the-art results on the ADL-X benchmark suite and outperforming strong baselines on our proposed benchmark. Code, models, and data will be released at https://github.com/dominickrei/EgoExo4ADL.

1 Introduction

Vision Language Models (VLMs) [3, 44, 64, 70] have achieved strong performance across a wide range of multi-modal understanding tasks, from open-ended video question answering [33, 68] to complex spatial reasoning [24, 50]. Existing VLMs for video understanding work by coupling Large Language Models (LLMs) [40, 47] together with pretrained vision encoders [48, 69] to enable powerful cross-modal reasoning capabilities. In practice, these models are primarily

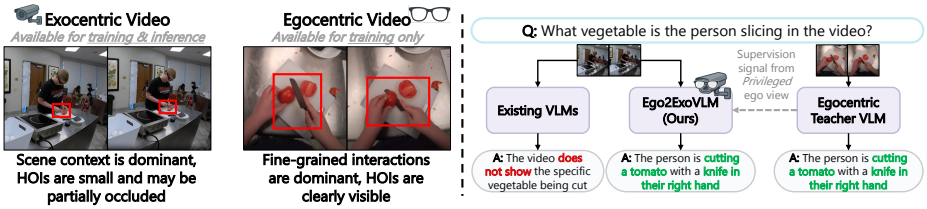

Fig. 1: Left: Egocentric and exocentric viewpoints emphasize different visual cues. In exocentric views, the hand-object interactions defining the action appears small or partially occluded, while egocentric views reveal these cues clearly. Red boxes highlight the interaction region. **Right:** Existing VLMs observing exocentric video struggle to capture egocentric properties. Ego2ExoVLM leverages ego-exo video pairs during training, where the egocentric view acts as *privileged* supervision available only during training, enabling VLMs to recover egocentric properties from exocentric video.

trained on large-scale video-instruction datasets consisting of both egocentric and exocentric viewpoints [6,36,74]. Applying such VLMs to recognize Activities of Daily Living (ADL) from fixed monitoring cameras is a particularly promising direction, with applications in healthcare (e.g., monitoring the elderly, assessing cognitive decline) and assistive robotics.

However, these models struggle in ADL scenarios where the activities of interest are largely defined by egocentric properties, i.e., *fine-grained scene content* and *human-object interactions* (HOI) [16,17,54]. Egocentric and exocentric viewpoints naturally emphasize different visual cues: egocentric views expose fine-grained hand-object interactions and manipulation details, while exocentric views emphasize scene layout, body motion, and contextual information. Crucially, many egocentric interaction cues (e.g., which object is being manipulated) remain present but visually subtle in the exocentric view, making them difficult for models trained only on exocentric supervision to recover. As a result, models trained primarily on exocentric observations often fail to capture the egocentric properties that define many ADL. We identify two primary causes: (i) existing VLMs [51] are trained on instruction-video pairs derived from exocentric ADL datasets [12, 21, 32] that emphasize high-level scene context, full-body motion, and environmental details, thereby overlooking egocentric properties that are less visually salient; and (ii) they rarely exploit the corresponding egocentric videos at inference, in which egocentric properties are more directly observable. This is illustrated in Figure 1, where existing VLMs observing exo videos fail to capture egocentric properties, while ego-observing VLMs readily capture these details. However, these egocentric observations are difficult to obtain in real-world ADL scenarios, where wearable cameras are impractical to deploy. This motivates a key question: *how can we train VLMs to infer egocentric properties from exocentric video inputs?*

Consequently, we present **Ego2ExoVLM**, a novel framework that leverages time-synchronized ego-exo video pairs to transfer egocentric understanding to models observing exocentric videos. Unlike standard instruction tuning [31] or distillation [46], where both teacher and student observe the same video, our

approach leverages a *privileged* egocentric viewpoint that exposes fine-grained interaction cues which are present in the scene but not visually prominent from the exocentric view. By learning from this privileged supervision during training, Ego2ExoVLM encourages VLMs operating on exocentric inputs to recover egocentric properties that would otherwise remain latent. At inference, Ego2ExoVLM takes *only* exocentric video as input and generates outputs that emulate the ego-observing teacher, which naturally captures the egocentric properties we desire to learn from exocentric videos.

Ego2ExoVLM introduces two complementary components to facilitate ego-to-exo knowledge transfer. (i) **Ego2Exo Sequence Distillation**: we transfer egocentric reasoning to exocentric observations through language-level distillation, encouraging the model to generate responses consistent with those produced from the privileged egocentric viewpoint. However, response consistency alone is insufficient due to the large ego-exo viewpoint gap, where many interaction cues occupy only a small portion of the exocentric view. (ii) **Ego Adaptive Visual Tokens**: to help the model recover these cues we take inspiration from *visual probing* approaches [52], introducing a set of learnable tokens that interact with intermediate visual representations and encourage the extraction of discriminative spatio-temporal regions corresponding to egocentric interactions. Together, these components enable **Ego2ExoVLM** to leverage privileged egocentric supervision during training to recover egocentric properties from purely exocentric inputs. Further, to evaluate this capability, we introduce **Ego-in-Exo Perception**, a benchmark designed to measure egocentric understanding from exocentric videos. The benchmark leverages time-synchronized ego-exo video pairs where questions are derived from egocentric observations while evaluation is performed using only the paired exocentric videos.

To the best of our knowledge, Ego2ExoVLM is the first VLM that explicitly learns exocentric representations from privileged egocentric supervision. While the inverse direction, i.e., exo-to-ego transfer, has been explored, works exploring ego-to-exo transfer remain limited, particularly within the VLM domain. Importantly, the two directions address fundamentally different cues: exo→ego transfer primarily propagates coarse contextual information such as scene layout and body motion, whereas ego→exo transfer requires recovering subtle interaction cues such as hand-object interactions that are often visually ambiguous from exocentric viewpoints and therefore difficult to learn through exocentric supervision alone. Prior approaches for cross-view knowledge transfer are designed for vision-language encoders (e.g., CLIP [48]), and mostly rely on feature-level knowledge transfer in the visual space [35, 42, 46]. Interestingly, we find that such feature-level knowledge transfer is ineffective in VLMs (Section 7.3); instead, we find that *language* serves as a better medium to bridge the knowledge gap between viewpoints. Some recent works explore cross-view transfer without time-synchronized ego-exo data by relying on pseudo-paired videos [42, 61] or synthetically generated ego viewpoints [13]. While such strategies can expand training data, synthetic ego generation is often impractical in ADL settings where activities are captured by fixed monitoring cameras and human-object

interactions appear at low spatial resolutions. We therefore focus on the realistic ADL scenario with time-synced ego-exo data; however, we later demonstrate (in Section 8) that Ego2ExoVLM remains effective even when trained with non-synchronized ego-exo supervision. We summarize our contributions as follows:

- We introduce **Ego2ExoVLM**, the first VLM framework trained with time-synchronized ego-exo video pairs to infer egocentric properties within exocentric inputs, addressing the underexplored problem of ego→exo transfer for ADL understanding.
- We propose two complementary components for enabling cross-view transfer in VLMs: **Ego2Exo Sequence Distillation**, that transfers egocentric reasoning through language-level sequence distillation from a privileged egocentric viewpoint, and **Ego Adaptive Visual Tokens**, a set of learnable tokens to surface ego-relevant spatio-temporal cues from exocentric videos.
- We introduce **Ego-in-Exo Perception**, a benchmark of 3.9k MCQs designed to evaluate egocentric understanding from exocentric videos, and validate Ego2ExoVLM across this benchmark and the ADL-X benchmark suite.

2 Related Work

Vision Language Models for Video Advancements in Large Language Models [4, 8, 56] and large-scale video-text datasets [26, 63, 73, 74] have led to Vision Language Models (VLMs) [23, 25, 29, 37, 71, 73] with impressive video understanding capabilities. Most existing VLMs are trained on datasets containing both egocentric and exocentric videos [15, 74], allowing them to exhibit egocentric understanding only when the input itself is egocentric. However, these models fail to demonstrate egocentric reasoning when observing the same activities from an exocentric viewpoint. Our work is the first to explicitly address this limitation by developing a VLM that transfers egocentric understanding to exocentric videos through paired, time-synchronized ego-exo supervision.

Ego-Exo Video Representation Learning. Various prior works have explored video understanding from ego-exo views, and can be categorized [55] into joint-learning and view transfer approaches. Joint learning approaches [42, 53, 61, 62, 65, 67] aim to learn a unified representation space for both views. For example, Actor and Observer [53] trains a dual-stream CNN to contrastively align ego and exo features, while AE2 [65] uses temporal alignment as a contrastive learning objective. In real-world scenarios where only a single view is available for inference, view transfer approaches [2, 28, 46, 49, 58, 60] aim to leverage knowledge from one view to enhance understanding of the other. For example, Ego-Exo [28] uses egocentric auxiliary tasks to pre-train a 3D-CNN to learn egocentric properties from exocentric videos, Quattrocchi *et al.* [46] distills the visual features from an exo-trained teacher to an ego student, EMBED [13] trains an encoder with synthetic ego data using hand crops derived from exo videos, and ViewpointRosetta [35] uses diffusion models to learn a mapping between the visual feature spaces of ego and exo. Recent cross-view translation methods [34, 38, 45] are also related,

but they aim to synthesize or structurally reconstruct observations across viewpoints, whereas our method targets understanding rather than pixel generation. Most of these approaches focus exclusively on transferring knowledge from the exocentric to egocentric viewpoint, and are not designed for VLMs. In contrast, our work investigates the opposite direction of transferring knowledge from the egocentric to exocentric viewpoint in VLMs.

3 VLM Preliminaries

The standard VLM is composed of a vision encoder $E(\cdot)$, a vision-language connector $\phi(\cdot)$, and a language model $LM(\cdot)$. Given a video $\boldsymbol{V}$, natural language query $\mathbf{Q}$, and a corresponding answer $\mathbf{A}$, the vision encoder first extracts frame-level spatio-temporal tokens from the video

$$\mathbf{X} = E(\boldsymbol{V}) \in \mathbb{R}^{T \times N \times d_v},$$

where T is the number of frames sampled from the video, N is the number of spatial patches, and d_v is the embedding dimension of the vision encoder. These tokens are then mapped to the language model's embedding space, d_{lm}, using the vision-language connector:

$$\mathbf{Z} = \phi(\mathbf{X}) \in \mathbb{R}^{T \times N \times d_{\text{lm}}},$$

Conditioned on these projected visual tokens, the training objective of the VLM is to auto-regressively maximize the likelihood of the answer given the video and query

$$\mathcal{L}_{\text{VLM}} = -\sum_{m=1}^{M} \log P(\mathbf{a}_m \mid \mathbf{A}_{<m}, \mathbf{Q}, \mathbf{Z}), \tag{1}$$

where M is the number of tokens in the answer, and $\mathbf{A}_{<m} = (\mathbf{a}_1, \ldots, \mathbf{a}_{m-1})$ denotes the sequence of tokens prior to the auto-regressive decoding step m.

4 Task Formulation

Given time-synchronized pairs of ego-exo activity videos $(\boldsymbol{V}_{\text{ego}}, \boldsymbol{V}_{\text{exo}})$ along with a natural language query $\mathbf{Q}$, we aim to learn a VLM that can accurately respond to the query using only the exocentric video at inference time. Note that the answer $\mathbf{A}$ is not explicitly given at training time. Formally, during training, we have access to triplets $(\boldsymbol{V}_{\text{ego}}, \boldsymbol{V}_{\text{exo}}, \mathbf{Q})$, but at test time, the model only has access to $(\boldsymbol{V}_{\text{exo}}, \mathbf{Q})$. The challenge of this task lies in the fact that the language queries target egocentric properties of the video, which are not salient from the exocentric viewpoint. The necessity of this task is grounded in the understanding of Activities of Daily Living, where such egocentric properties define the activities of interest; however, the egocentric view is rarely available in real-world ADL scenarios. To solve this, we leverage time-synchronized ego-exo video pairs during training to transfer egocentric understanding from an ego-observing teacher VLM to an exo-observing student VLM.

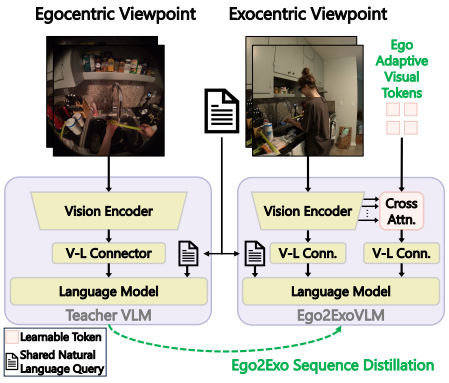

Fig. 2: Overview of our proposed Ego2ExoVLM. A teacher VLM observing egocentric viewpoint videos transfers knowledge to a student VLM through Ego2Exo Sequence Distillation. The student VLM is augmented with Ego Adaptive Visual Tokens, which are cross-attended with exocentric visual features.

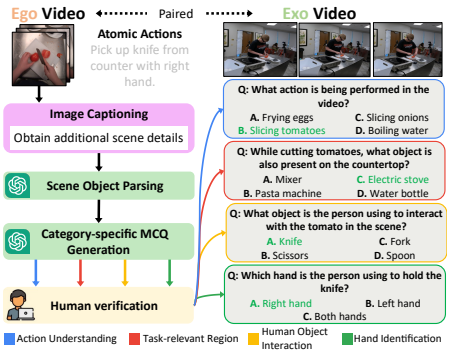

Fig. 3: Ego-in-Exo Perception data curation pipeline. The egocentric viewpoint is used to generate question-answer pairs for the corresponding time-synched exocentric viewpoint. In other words, at inference we query the exocentric video with egocentric questions. Additional examples available in the Appendix.

5 Proposed Method

This section introduces our proposed method, **Ego2ExoVLM**, which addresses ego-to-exo viewpoint transfer through knowledge distillation. Ego2ExoVLM training follows a teacher-student paradigm (see Figure 2) and comprises two complementary components: (1) **Ego2Exo Sequence Distillation** that transfers egocentric semantics from an egocentric teacher to an exocentric student, and (2) **Ego Adaptive Visual Tokens** that enhances the student's ability to acquire egocentric semantics from the teacher.

Both the teacher and student are VLMs sharing the architecture described in Section 3. The **teacher VLM** $\mathcal{T}$ processes the egocentric video to generate an answer that implicitly captures egocentric properties, which then serves as a source of egocentric supervision for the exocentric **student VLM** $\mathcal{S}$, which processes the corresponding exocentric video. During training, the teacher remains frozen and serves as a high-level semantic bridge between the two viewpoints.

5.1 Ego2Exo Sequence Distillation

Our goal is to leverage the time-synchronized ego-exo videos $(\boldsymbol{V}_{\text{ego}}, \boldsymbol{V}_{\text{exo}})$ to transfer egocentric understanding from $\mathcal{T}$ to $\mathcal{S}$. In Ego2Exo Sequence Distillation, the teacher observes the activity from the egocentric viewpoint to generate a language sequence capturing the egocentric properties of the activity, while the student observes the activity from the exocentric viewpoint and is supervised by the teacher to reproduce its outputs. Formally, the teacher VLM processes the egocentric video and query to generate the language sequence

$\mathbf{A}_{ego} = \mathcal{T}(\boldsymbol{V}_{ego}, \mathbf{Q})$, which serves as a pseudo-ground truth answer to guide the student VLM. During training, the objective of the student is to auto-regressively maximize the likelihood of $\mathbf{A}_{ego}$

$$\mathcal{L}_{\text{VLM}} = -\sum_{m=1}^{M} \log P(\mathbf{a}_m^{ego} \mid \mathbf{A}_{<m}^{ego}, \mathbf{Q}, \mathbf{Z}^{exo}) \tag{2}$$

This objective enables cross-view knowledge transfer by enforcing consistency between the egocentric teacher, which implicitly understands the egocentric properties of the video, and the exocentric student.

5.2 Ego Adaptive Visual Tokens

While the sequence-level distillation described above provides the supervisory signal necessary to transfer viewpoint knowledge from teacher to student, we observe that exocentric videos contain many visual distractors (e.g., background content) that dilute the spatial cues relevant for egocentric understanding, limiting the effectiveness of cross-view transfer. To enhance the effectiveness of this transfer, we introduce a set of learnable Ego Adaptive Visual Tokens that guide the student VLM to focus on spatial regions carrying ego-relevant cues. These additional tokens enable the student to selectively extract exocentric visual features that better align with the egocentric language sequence. In essence, the ego adaptive tokens are supervised through the egocentric signal to capture spatio-temporal patterns that are less prominent in the exocentric view.

Concretely, the learnable tokens are implemented as *visual probes* [52] that are cross-attended with the intermediate representations of the vision encoder, allowing them to leverage both low- and high-level pixel information to facilitate the extraction of discriminative exocentric representations for bridging the gap between the two viewpoints. Letting the encoder output features at layer ℓ be $\mathbf{X}^{\ell} \in \mathbb{R}^{T \times N \times d_v}$ and the learnable ego adaptive tokens be $\mathbf{E}^{\ell} \in \mathbb{R}^{K \times d_v}$, cross-attention is applied as follows

$$\tilde{\mathbf{E}}^{\ell} = \text{softmax}\left(\frac{(\mathbf{E}^{\ell}\boldsymbol{W}_q^{\ell})(\mathbf{X}^{\ell}\boldsymbol{W}_k^{\ell})^{\top}}{\sqrt{d_v}}\right)(\mathbf{X}^{\ell}\boldsymbol{W}_v^{\ell}),$$

where $(\boldsymbol{W}_q^{\ell}, \boldsymbol{W}_k^{\ell}, \boldsymbol{W}_v^{\ell})$ are layer-specific projection matrices. The tokens after the final layer (denoted $\tilde{\mathbf{E}}$ for simplicity) are projected into the language model's embedding space through a dedicated connector $\phi_{\text{ego}}(\cdot)$ and concatenated with the exocentric visual embeddings before being fed to the language model. Thus, the training objective of Ego2ExoVLM is defined as

$$\mathcal{L}_{\text{VLM}} = -\sum_{m=1}^{M} \log P(\mathbf{a}_m^{ego} \mid \mathbf{A}_{<m}^{ego}, \mathbf{Q}, \mathbf{Z}^{exo}, \phi_{\text{ego}}(\tilde{\mathbf{E}})) \tag{3}$$

With this objective, the teacher is able to transfer egocentric cues to the student by leveraging the *shared content* of time-synchronized ego-exo videos.

Table 1: Comparison with existing benchmarks. Columns indicate benchmark properties: **Ego-Exo**: contains paired ego-exo videos, **Time-sync**: videos are time-synchronized, **Video-QA**: suitable for evaluating VLMs, **Ego-curated QA**: video-QA pairs are curated only from the egocentric viewpoint.

Benchmark	Ego-Exo?	Time-sync?	Video-QA?	Ego-curated QA?
Kinetics [5]	✗	✗	✗	✗
VideoMME [14]	✗	✗	✓	✗
EgoMCQ [30]	✗	✗	✗	✓
EgoSchema [39]	✗	✗	✓	✓
EgoMemoria [66]	✗	✗	✓	✓
LEMMA [22]	✓	✓	✗	✗
EgoExoLearn [20]	✓	✓	✗	✗
EgoExoBench [18]	✓	✓	✓	✗
Ego-in-Exo Perception	✓	✓	✓	✓

Table 2: State-of-the-art comparison on the Ego-in-Exo Perception benchmark. Action Und.: *Action Understanding*, Task Regions: *Task Relevant Regions*, HOI: *Human Object Interactions*, Hand Ident.: *Hand Identification*.

Method	Action Und.	Task Regions	HOI	Hand Ident.	Avg.
Evaluate on ego viewpoint (upper bound)					
Teacher ($\mathcal{T}$)	81.8	83.9	78.9	67.1	77.9
Evaluate on exo viewpoint					
LLAVIDAL [51]	15.2	33.3	31.0	53.8	33.3
VideoLLaMA2 [7]	35.3	42.8	34.6	48.4	40.3
Qwen2-VL [57]	38.5	48.9	37.4	56.4	48.8
Qwen2.5-VL [47]	68.2	75.9	70.8	64.8	69.9
VideoLLaMA3 [70]	66.5	79.6	74.5	64.1	71.2
EE4D-VideoLLaMA3	59.5	66.6	71.6	65.0	65.7
Ego2ExoVLM (Ours)	**73.4**	**82.0**	**76.2**	**65.3**	**74.2**

6 Ego-in-Exo Perception Benchmark

In this section, we introduce Ego-in-Exo Perception, a benchmark containing a total of 3,881 multiple choice questions (MCQs) designed to evaluate egocentric understanding from exocentric video inputs. It consists of time-synchronized videos of skilled human activities captured simultaneously from both egocentric and exocentric viewpoints. We propose this benchmark to answer the single question: *How well do VLMs capture egocentric properties from exocentric videos?*

The key novelty of Ego-in-Exo Perception lies in how it is designed to answer this question. We posit that to measure egocentric understanding from exocentric videos, the question-answer (QA) pairs used in our benchmark should be curated using only the corresponding egocentric view. Our intuition behind this is that if QAs are curated based on what is salient in ego, then answering these questions from the exo video alone directly measures the model's ability to capture ego properties from exo videos. Samples from our benchmark are shown in Figure 3 (right), with more samples presented in the supplementary materials.

6.1 Data Curation

We curate Ego-in-Exo Perception from the EgoExo4D [16] dataset, a large-scale dataset of long videos of skilled human activities captured from ego and exo viewpoints. We only consider videos from the keystep recognition *test* subset of EgoExo4D, and only keep videos depicting "*cooking*" activities, as these activities contain a high density and diversity of human-object interactions [11], making them an ideal testbed for measuring the egocentric properties we are interested in. This filtered subset provides us with 2,319 unique videos along with 10K human annotations detailing the atomic actions within each video. Our benchmark contains four tasks that evaluate different aspects of egocentric understanding: Action Understanding, Task-relevant Region, Human-Object Interactions, and Hand Identification.

The videos and annotations are then processed through a multi-stage pipeline, as illustrated in Figure 3, to obtain MCQs. We first extract a list of the objects

visible in the egocentric video by applying an image captioning model [19] at 5fps and parsing its output using an LLM [43] to obtain a list of scene objects. Next, we generate MCQs across the four tasks using an LLM [43] with category-specific prompts applied to the atomic action annotations, supplemented with the list of scene objects. We find that supplementing the annotations with this list enables the generation of more diverse question types with challenging negative distractors. Finally, we perform a comprehensive human verification with four participants to filter out hallucinated questions, trivial distractors, and questions that are unanswerable from the exocentric view due to occlusions or poor camera framing, yielding our final benchmark of 3,881 questions. A more detailed description of each step in this pipeline is provided in the Appendix.

6.2 Comparison with Existing Benchmarks

Table 1 compares Ego-in-Exo Perception with existing video understanding benchmarks. While several benchmarks exist to evaluate exocentric [5, 14] and egocentric [22, 30, 39, 66] understanding, they lack the time-synchronized ego-exo videos. Existing ego-exo benchmarks like LEMMA [22] and EgoExoLearn [20] provide time-synchronized ego-exo videos, but are not designed to evaluate VLMs. EgoExoBench [18] contains time-synchronized ego-exo videos and is designed for VLM evaluation, but uses both viewpoints in its data curation, making it unclear whether VLMs succeed by capturing egocentric properties or by exploiting information already salient in the exocentric view. In contrast, our benchmark uniquely leverages time-synchronized ego-exo videos, curating QA pairs from the egocentric viewpoint alone.

6.3 Evaluating Potential Bias in Ego-in-Exo Perception.

Table 3: Bias analysis on Ego-in-Exo Perception.

Method	Inputs	Action Und.	Task Regions	HOI	Hand Ident.	Avg.
GPT-4.1-Mini	Text only	56.4	56.3	55.3	54.4	55.7
	Text + Image	71.4	79.8	70.3	55.7	69.8
	Text + Video	**77.5**	76.6	75.9	59.5	73.4
GPT-5.1	Text + Video	74.6	**84.8**	**77.2**	54.8	72.9
Ego2ExoVLM	Text + Video	73.4	82.0	76.2	**65.3**	**74.2**

To assess whether Ego-in-Exo Perception can be solved via linguistic priors or simple visual shortcut biases, we evaluate GPT-4.1-Mini and GPT-5.1 under three input settings: *Text only* (question without video), *Text + Image* (single center frame), and *Text + Video* (full video). As shown in Table 3, performance drops substantially in the Text-only setting (55.7% Avg.), indicating limited linguistic bias. Providing only the center frame improves results (69.8%), but still underperforms full video input (73.4%), suggesting that temporal information is necessary. Notably, Ego2ExoVLM achieves higher overall accuracy than GPT-4.1-Mini in the full Video setting (74.2% vs 73.4%), demonstrating that the benchmark rewards viewpoint-specific reasoning rather than scale alone and that the benchmark exhibits relatively weak biases.

7 Experiments

In Section 7.1, we detail the experimental setup. We then compare Ego2ExoVLM to the state-of-the-art on two activity understanding datasets in Section 7.2. Finally, in Section 7.3, we perform ablations on the Ego-in-Exo Perception benchmark and analyze the learned representations of Ego2ExoVLM.

Table 4: State-of-the-art comparison on the ADL-X benchmark. Methods are categorized by image captioning models paired with LLMs and vision language models. Legend for ADL Video Description tasks: [Cor: *Correctness of Information*, Do: *Detail Orientation*, Ctu: *Contextual Understanding*, Tu: *Temporal Understanding*, Con: *Consistency*]

Method	ADL MCQ					ADL-VD (Charades)						ADL-VD (Toyota Smarthome)					
	Charades	AR	SH	AR LEMMA	TC TSU	TC Avg	Cor	Do	Ctu	Tu	Con	Avg	Cor	Do	Ctu	Tu	Con	Avg
Image captioners + LLM																		
CogVLM [59] + GPT [4]	52.3	42.5	32.0	23.6	37.6	42.0	62.0	49.6	36.5	32.8	44.6	55.2	72.0	60.6	30.2	48.5	53.3	
CogVLM [59] + Llama [56]	52.8	43.2	32.5	22.5	37.8	40.2	61.8	49.5	36.5	33.5	44.3	49.8	66	56.6	29.8	40.2	48.5	
BLIP2 [27] + GPT [4]	50.2	39.6	28.9	20.2	34.7	39.8	60.2	47.8	36.0	37.2	44.2	48.8	66.6	63.6	45.6	39.8	52.9	
Vision Language Models																		
VideoChatGPT [37]	51.0	39.6	31.4	20.9	35.7	26.1	45.2	35.6	21.4	31.2	31.9	31.2	52.8	78.2	64.8	45.6	54.5	
VideoLLaMA [71]	40.2	44.8	32.6	24.6	35.6	22.2	42.5	33.8	20.2	34.5	30.6	57.8	62.0	62.4	48.2	44.4	54.9	
VideoLLaVA [29]	41.8	49.2	30.0	25.5	36.6	23.6	46.4	34	20.6	33.5	31.6	30.8	54.8	42.4	30.4	44.5	40.6	
Chat-UniVi [23]	53.1	48.1	32.3	36.4	42.5	36.5	54.5	46.6	32.2	35.9	41.1	56.8	66.9	79.0	50.0	56.6	61.9	
ADL-X-ChatGPT [51]	51.0	44.5	28.6	29.5	38.4	40.6	50.6	49.8	30.6	40.2	42.4	62.4	79.4	70.8	51.2	60.4	64.8	
LLAVIDAL [51]	55.2	48.1	34.3	38.2	44.0	45.8	64.2	57.0	36.4	39.4	48.6	66.0	86.2	79.6	50.0	72.4	70.8	
VideoLLaMA3 [70]	92.0	70.6	66.6	78.3	76.9	73.5	73.7	75.8	68.6	61.6	70.6	88.1	91.8	92.6	79.2	**84.4**	87.2	
EE4D-VideoLLaMA3	90.9	71.3	60.2	77.7	75.0	80.4	78.8	82.0	73.8	61.9	75.4	**91.5**	92.5	95.8	79.0	73.5	86.4	
Ego2ExoVLM (Ours)	**93.1**	**72.0**	**67.4**	**82.6**	**78.8**	**80.5**	**78.8**	**82.3**	**73.8**	**62.4**	**75.5**	90.8	**93.1**	**95.9**	**80.8**	82.9	**88.7**	

7.1 Experimental Setting

Teacher and Student VLMs. Both the teacher and student adopt identical VLM architectures consisting of a SigLIP [69] vision encoder (embedding dim. $d_v = 1152$) and a Qwen2.5 [47] language model (embedding dim. $d_{\text{lm}} = 3584$), initialized from the pretrained weights of VideoLLaMA3 [70]. During training, the teacher VLM remains frozen and the following components of the student VLM remain trainable: vision-language connector, cross-attention projection matrices (for ego-adaptive tokens), and low-rank adaptation (LoRA with rank $r = 64$) weights in the LLM are updated. The vision encoder remains frozen in all experiments. Training is conducted on 4 NVIDIA H200 GPUs for 3 epochs, with a learning rate of 1×10^{-5} for the LoRA parameters and 2×10^{-6} for all other parameters.

Student Training Data. The student VLM is trained on a total of 9k ego-exo videos from the keystep recognition training set of EgoExo4D [16]. For each ego-exo video pair, we generate three distinct natural language queries (i.e., **Q**) by first sampling one query from each of three predefined categories: descriptive, temporal, and spatial (provided in supplementary materials). We then process the sampled query along with the egocentric video with the frozen pre-trained teacher VLM which generates an answer (**A**). This query-answer pair then serves as supervision to the student VLM, yielding a total of 28k query-answer pairs.

Benchmarks. We evaluate Ego2ExoVLM across 10 tasks that rely on egocentric understanding from exocentric videos. The **Ego-in-Exo Perception**

benchmark, introduced in this work, is derived from the Ego-Exo4D [16] dataset and consists of 4 MCQ tasks introduced in Section 6: Action Understanding, Task-relevant Regions, Human Object Interactions, and Hand Identification. The **ADL-X** benchmark [51] evaluates VLM's understanding on multiple Activities of Daily Living (ADL) datasets [10,12,22,32]. Although ADL videos are recorded only from the exocentric viewpoint, the activities center around hand-object interactions and fine-grained motion details, making ADL a natural setting to assess egocentric understanding from exocentric videos. The benchmark contains 4 MCQ tasks (ADL-MCQ) and 2 video description tasks (ADL-VD). ADL-MCQ contains both Action Recognition (AR) and Temporal Completion (TC) tasks. Accuracy is reported for all MCQ tasks and Video-ChatGPT description metrics [37] are reported for description tasks.

7.2 Comparison with SoTA

Ego-in-Exo Perception Benchmark. Table 2 compares Ego2ExoVLM with existing VLMs on our proposed Ego-in-Exo Perception benchmark. We include the teacher model's performance on egocentric videos as an upper bound (77.9% average accuracy), representing the level of understanding achievable with direct access to the egocentric viewpoint. We also present a variant of VideoLLaMA3 finetuned on the exocentric videos of Ego-Exo4D (EE4D-VideoLLaMA3). LLAVIDAL [51] struggles significantly (33.3%), particularly on Action Understanding (15.2%) despite being trained on ADL data, highlighting that standard exocentric training does not capture the egocentric properties targeted by our benchmark. VideoLLaMA3 [70], whose weights are used to initialize our student VLM, achieves higher accuracy than other baselines, likely due to its large-scale pretraining data. However, Ego2ExoVLM outperforms this strong baseline by +3.0 % (71.2% vs 74.2%). Comparing Ego2ExoVLM with EE4D-VideoLLaMA3, our method achieves superior performance (74.2% vs 65.7%), demonstrating that leveraging time-synchronized ego-exo pairs for cross-view knowledge transfer is more effective than training solely on exo videos from the same dataset.

ADL-X. Table 4 compares Ego2ExoVLM against state-of-the-art methods on the ADL-X benchmark. Image captioning models such as CogVLM [59] achieve moderate performance on ADL tasks even when paired with large language models like GPT-4 (44.6% on Charades Description compared to 75.5% on Ego2ExoVLM). Among existing VLMs, pre-trained VideoLLaMA3 performs better on ADL tasks, and EE4D-VideoLLaMA3 can further improve performance by leveraging in-domain data. However, our approach surpasses both, reaching 78.8% on ADL-MCQ, 75.5% on Charades Descriptions, and 88.7% on Toyota Smarthome Descriptions. Ego2ExoVLM also outperforms LLAVIDAL [51], a VLM specifically tailored for ADL understanding, demonstrating Ego2ExoVLM's understanding of the egocentric properties relevant in ADL.

7.3 Ablation and Analysis

Ablating Components. Table 5 ablates each component in Ego2ExoVLM. Crucially, simply exposing the model to both ego and exo videos (Ego+Exo)

Table 5: Ablating Ego2ExoVLM components. Seq Dist. = *Ego2Exo Sequence Distillation*, Ego Tok. = *Ego Adaptive Visual Tokens*.

Training View	Seq Dist.	Ego Tok.	Action Und.	Task Regions	HOI	Hand Ident.	Avg.
Exo	✗	✗	59.5	66.6	71.6	65.0	65.7
Ego+Exo	✗	✗	59.2	68.5	71.1	65.0	66.0
Exo	✓	✗	67.2	67.9	73.8	65.3	68.5
Exo	✗	✓	70.0	79.4	74.3	65.1	72.2
Exo	✓	✓	**73.4**	**82.0**	**76.2**	**65.3**	**74.2**

Table 6: Alternative Ego Adaptive Visual Tokens Strategies.

Token Strategy	Action Und.	Task Regions	HOI	Hand Ident.	Avg.
No tokens	67.2	67.9	73.8	65.3	68.5
VE Self-attention	68.2	69.2	74.0	68.3	69.9
Pre-trained CA	**74.3**	79.9	74.1	65.1	73.4
Ego2ExoVLM	73.4	**82.0**	**76.2**	**65.3**	**74.2**

Table 7: Alternative Viewpoint Transfer Strategies.

Transfer Strategy	Action Und.	Task Regions	HOI	Hand Ident.	Avg.
Visual Feature Dist. [46]	61.2	77.7	68.0	64.2	67.8
LLAVIDAL-style [51]	60.2	66.1	70.3	65.0	65.4
Exo2EgoVLM [72]	65.9	75.1	63.8	81.1	70.3
EgoAda Tokens (Vis.)	69.5	79.4	74.1	65.3	72.1
EgoAda Tokens (Lang.)	69.2	78.2	72.0	64.9	71.1
Ego2ExoVLM	**73.4**	**82.0**	**76.2**	**65.3**	**74.2**

yields negligible gains (+0.3%). This demonstrates that viewpoint-conditioned semantics are not learned through shared training data alone and require explicit cross-view supervision. In contrast, training on exo videos with sequence distillation alone improves average performance to 68.5%. Ego-adaptive tokens alone yield stronger improvements at 72.2%, particularly on the HOI task, indicating their effectiveness to enhance ego understanding. Combining both components achieves the best overall performance (74.2% average), with consistent gains across all tasks, demonstrating the complementary nature of Ego2Exo Sequence Distillation and Ego Adaptive Visual Tokens. These results demonstrate that both modules in are instrumental to enable ego-to-exo knowledge transfer.

Alternate Knowledge Transfer Strategies. Table 7 compares our Ego2Exo Sequence Distillation against alternative viewpoint transfer strategies. In the *Visual Feature Dist.* baseline, we train a VLM on egocentric videos and apply an MSE distillation loss between the pooled visual features of the frozen ego teacher and the exo student, analogous to the viewpoint transfer approach in Quattrocchi *et al.* [46]. In LLAVIDAL-style baseline, we train a VLM with ego-adaptive tokens on ego videos and directly pass these tokens as additional input to the language model when training the exo student, similar to LLAVIDAL [51]. In EgoAda Tokens, we explore viewpoint transfer using the ego adaptive visual tokens themselves: we train an ego teacher VLM with additional tokens and perform feature-level distillation between the frozen ego tokens and exo tokens either after the visual encoder (Vis.), or after the language model (Lang.). We find that our proposed strategy surpasses all others, including the visual feature strategies used in prior non-VLM-based approaches (67.8% vs 74.2%), demonstrating the superior effectiveness of language-level knowledge transfer in VLMs. Notably, Exo2EgoVLM [72], designed for exo→ego transfer, underperforms in the ego→exo setting (70.3% vs 74.2%), indicating that methods effective for exo2ego do not directly translate to ego2exo transfer.

Alternative Ego Adaptive Visual Tokens Strategies. We explore alternative integration strategies for ego adaptive tokens in Table 6. In VE Self-attention, the cross-attention matrices are removed and the tokens are directly appended to the visual features, thus they only interact with visual features through the self-attention of the vision encoder. In Pre-trained CA, we first train a VLM with ego-adaptive tokens on ego videos and initialize the student with the parameters of this model. VE Self-attention achieves modest gains compared to the VLM without tokens, but the lack of cross-attention limits the ability to learn robust token representations. While Pre-trained CA approaches Ego2ExoVLM's performance, we find that it is less effective.

Analysis on number of Ego Adaptive Visual Tokens. In Figure 4, we present the accuracy of Ego2ExoVLM with varying numbers of Ego Adaptive Visual Tokenson Ego-in-Exo Perception. We find that our model's sensitivity to the number of tokens is low, and set the default to a value of 16 tokens for all our experiments as a reasonable trade-off between efficiency and performance.

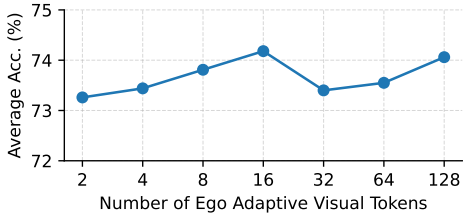

Fig. 4: Ablation on number of Ego Adaptive Visual Tokens. Average accuracy on Ego-in-Exo Perception is reported.

Attention Visualization. In Figure 5, we visualize the attentions of the visual encoder of an untrained student VLM, and of our Ego Adaptive Visual Tokens's to a video frames from two different videos. For the vision encoder, we visualize attention using attention rollout [1], for our ego adaptive tokens, we visualize their attentions averaged across all tokens. We observe that the VLM vision encoder attends broadly to scene-level content, while our Ego Adaptive Visual Tokens focus on localized regions where relevant interactions occur, such as human object interactions.

Qualitative Results on ADL-X. Figure 6 shows qualitative results of three VLMs: VideoLLaMA3, EE4D-VideoLLaMA3, and our proposed Ego2ExoVLM. We present example outputs of these VLMs on the Smarthome-AR, TSU-TC, and Charades Descriptions subsets of ADL-X. These results demonstrate that Ego2ExoVLM accurately identifies specific actions, object, and spatial details that are critical for ADL understanding.

Table 8: Ego2ExoVLM integrated into a different VLM backbone.

Model	Backbone	Action Und.	Task Regions	HOI	Hand Ident.	Avg
Baseline	CLIP / Mistral	30.1	32.6	30.6	54.6	37.0
Ego2ExoVLM	CLIP / Mistral	**41.2**	**43.8**	**40.9**	**56.8**	**45.7**

Robustness to Alternative VLM Backbones. We investigate whether the improvements of Ego2ExoVLM generalize across different VLM backbones. Table 8 reports results using a CLIP-Large vision encoder and Mistral v0.2 language model, which represent a substantially weaker backbone compared to the SigLIP / Qwen2.5 backbone used in our main experiments. As expected, the absolute performance is lower due to the reduced model capacity and training data of the CLIP/Mistral backbone. Nevertheless, Ego2ExoVLM

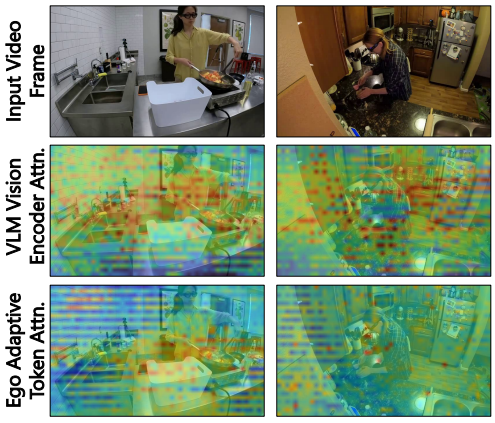

Fig. 5: Visualization of VLM Attention. We compare the visual encoder's attention to attention of the ego-adaptive tokens.

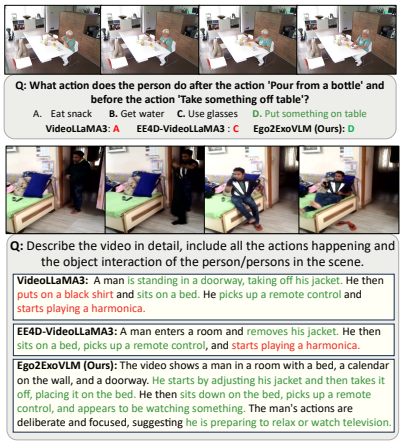

Fig. 6: Qualitative Results on the ADL-X dataset. Additional qualitative results available in the Appendix.

consistently improves over the baseline across all tasks, demonstrating that the gains from ego-to-exo supervision are not specific to our backbone of choice.

Table 9: Perplexity and KL-Divergence with Teacher VLM.

Method	Perplexity (↓)	KL-Divergence (↓)
Untrained Student	1.55	0.16
Ego2ExoVLM	**1.41**	**0.05**

Table 10: Ego2ExoVLM's effectiveness on pseudo-paired Ego4D and HowTo100M.

Model	Training Data	Action Und.	Task Regions	HOI	Hand Ident.	Avg
Baseline	Ego4D + HT100M	65.3	66.8	**75.1**	**65.3**	68.1
Ego2ExoVLM	Ego4D + HT100M	**69.5**	**77.6**	71.7	62.8	**70.4**

Perplexity and KL-Divergence with Teacher VLM. In Table 9, we compute perplexity and KL-Divergence [9] between the language outputs of the ego-observing teacher, an untrained student VLM, and Ego2ExoVLM on a subset of Ego-in-Exo Perception. Perplexity measures how closely a model's language output aligns with the teacher's, while KL-divergence measures the similarity between the logit-level outputs of each model. Ego2ExoVLM achieves lower scores on both metrics, indicating that the outputs of our Ego2ExoVLM are more closely aligned with those of the ego-observing teacher.

8 Learning from Unpaired Ego-Exo Videos.

We next test whether exact temporal alignment is necessary. Instead of time-synced video pairs, we construct pseudo ego-exo pairs by retrieving exocentric clips from HowTo100M [41] for each egocentric clip from Ego4D [15] using the same text-based retrieval process as in [61]. These pairs originate from different videos and datasets, with no shared frame-level correspondence. Table 10 shows that training on these unpaired but semantically matched videos still improves over the baseline (Avg 70.4% vs 68.1%). This experiment demonstrates that Ego2ExoVLM 's language-mediated cross-view supervision does not require strictly time synchronized ego-exo video pairs.

9 Conclusion

We addressed the challenge of understanding egocentric properties from exocentric video observations, which remains difficult for existing VLMs despite their strong performance on general video understanding tasks. We presented Ego2ExoVLM, a VLM trained to infer egocentric properties from exocentric videos through time-synchronized cross-view knowledge transfer. To the best of our knowledge, Ego2ExoVLM is the first VLM to transfer knowledge from the egocentric to the exocentric viewpoint. Our approach combines Ego2Exo Sequence Distillation, which transfers egocentric reasoning through language-sequence supervision, with Ego-Adaptive Visual Tokens that enhance attention to ego-relevant spatial regions. To evaluate this capability, we introduced Ego-in-Exo Perception, a benchmark of 3.9K questions curated from the egocentric viewpoint while evaluated using exocentric videos. Through comprehensive evaluation, we demonstrate state-of-the-art performance on the ADL-X benchmark suite and strong improvements over baselines on Ego-in-Exo Perception, showing that egocentric supervision can effectively enable exocentric models to recover egocentric interaction cues. We further show that Ego2ExoVLM remains effective even when trained with non-time-synchronized ego-exo data, demonstrating that strict temporal alignment is not required for effective cross-view transfer.

10 Acknowledgments

This work was supported in part by the National Science Foundation (IIS-2245652) and the University of North Carolina at Charlotte. Computational resources were provided by the NSF National AI Research Resource Pilot (NAIRR-240338) and the NCShare initiative. We would like to thank Michael Ryoo and Razvan Bunescu for their valuable insights and discussions.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Annual Meeting of the Association for Computational Linguistics (2020) 13
2. Ardeshir, S., Borji, A.: An exocentric look at egocentric actions and vice versa. In: Computer Vision and Image Understanding. vol. 171, pp. 61–68 (2018) 4
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 1
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 4, 10

5. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about kinetics-600. CoRR **abs/1808.01340** (2018), http://arxiv.org/abs/1808.01340 8, 9
6. Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Lin, B., Tang, Z., Yuan, L., Qiao, Y., Lin, D., Zhao, F., Wang, J.: Sharegpt4video: Improving video understanding and generation with better captions. In: Advances in Neural Information Processing Systems (2024) 2
7. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., Bing, L.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024), https://arxiv.org/abs/2406.07476 8
8. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/ 4
9. Cover, T.M., Thomas, J.A.: Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, USA (2006) 14
10. Dai, R., Das, S., Sharma, S., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022). https://doi.org/10.1109/TPAMI.2022.3169976 11
11. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: European Conference on Computer Vision (ECCV) (2018) 8
12. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome: Real-world activities of daily living. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 833–842 (2019) 2, 11
13. Dou, Z.Y., Yang, X., Nagarajan, T., Wang, H., Huang, J., Peng, N., Kitani, K., Chu, F.J.: Unlocking exocentric video-language data for egocentric video representation learning (2024), https://arxiv.org/abs/2408.03567 3, 4
14. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Shan, C., He, R., Sun, X.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal large language models in video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 8, 9
15. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Ruiz, P., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A.,

Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18995–19012 (June 2022) 4, 14

16. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: 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 (2025) 2, 8, 10, 11
17. He, Y., Huang, Y., Chen, G., Lu, L., Pei, B., Xu, J., Lu, T., Sato, Y.: Bridging perspectives: A survey on cross-view collaborative intelligence with egocentric-exocentric vision. International Journal of Computer Vision **134**, 62 (2026) 2
18. He, Y., Huang, Y., Chen, G., Pei, B., Xu, J., Lu, T., Pang, J.: Egoexobench: A benchmark for first- and third-person view video understanding in mllms. arXiv preprint arXiv:2507.18342 (2025) 8, 9
19. Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., Tang, J.: Cogagent: A visual language model for gui agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 9
20. Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., Zhang, H., Dong, L., Wang, Y., Wang, L., Qiao, Y.: 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 (CVPR) (2024) 8, 9
21. Jang, J., Kim, D., Park, C., Jang, M., Lee, J., Kim, J.: ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2020) 2
22. Jia, B., Chen, Y., Huang, S., Zhu, Y., Zhu, S.C.: Lemma: A multiview dataset for learning multi-agent multi-view activities. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 8, 9, 11
23. Jin, P., Takanobu, R., Zhang, C., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 4, 10
24. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023) 1

25. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. ArXiv preprint arXiv:2408.03326 (2024) 4
26. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models (2024), https://arxiv.org/abs/2407.07895 4
27. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (2023), https://api.semanticscholar.org/CorpusID:256390509 10
28. Li, Y., Nagarajan, T., Xiong, B., Grauman, K.: 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 (CVPR). pp. 10995–11005 (2021) 4
29. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2024) 4, 10
30. Lin, K.Q., Wang, A.J., Soldan, M., Wray, M., Yan, R., Xu, E.Z., Gao, D., Tu, R., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. arXiv preprint arXiv:2206.01670 (2022) 8, 9
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 2
32. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 2, 11
33. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? In: European Conference on Computer Vision (2024) 1
34. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024) 4
35. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: Viewpoint rosetta stone: Unlocking unpaired ego-exo videos for view-invariant representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 3, 4
36. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Videogpt+: Integrating image and video encoders for enhanced video understanding (2024) 2
37. Maaz, M., Rasheed, H.A., Khan, S.H., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Annual Meeting of the Association for Computational Linguistics (ACL 2024) (2024) 4, 10, 11
38. Mahdi, M., Savov, N., Paudel, D.P., Gool, L.V.: From synchrony to sequence: Exo-to-ego generation via interpolation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2026) 4
39. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023), https://openreview.net/forum?id=JVlWseddak 8, 9
40. Meta: The llama 3 herd of models (2024) 1
41. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. arXiv preprint arXiv:1906.03327 (2019) 14

42. Ohkawa, T., Yagi, T., Nishimura, T., Furuta, R., Hashimoto, A., Ushiku, Y., Sato, Y.: Exo2egodvc: Dense video captioning of egocentric procedural activities using web instructional videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025) 3, 4
43. OpenAI: Gpt-4o system card (2024), https://arxiv.org/abs/2410.21276 9
44. OpenAI: Thinking with images (April 2025), https://openai.com/index/thinking-with-images/, accessed: June 27, 2026 1
45. Park, J., Ye, A.S., Kwon, T.: Egoworld: Translating exocentric view to egocentric view using rich exocentric observations. In: International Conference on Learning Representations (ICLR) (2026) 4
46. Quattrocchi, C., Furnari, A., Di Mauro, D., Giuffrida, M.V., Farinella, G.M.: Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs. In: European Conference on Computer Vision (ECCV) (2024) 2, 3, 4, 12
47. Qwen Team, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), https://arxiv.org/abs/2412.15115 1, 8, 10
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021) 1, 3
49. Rai, A., Buettner, K., Kovashka, A.: Strategies to leverage foundational model knowledge in object affordance grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1714–1723 (June 2024) 4
50. Ranasinghe, K., Shukla, S.N., Poursaeed, O., Ryoo, M.S., Lin, T.Y.: Learning to localize objects improves spatial reasoning in visual-llms. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12977–12987 (2024), https://api.semanticscholar.org/CorpusID:269043025 1
51. Reilly, D., Chakraborty, R., Sinha, A., Govind, M.K., Wang, P., Bremond, F., Xue, L., Das, S.: Llavidal: A large language-vision model for daily activities of living. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 2, 8, 10, 11, 12
52. Reilly, D., Govind, M.K., Xue, L., Das, S.: Viscop: Visual probing for video domain adaptation of vision language models. In: Proceedings of the European Conference on Computer Vision (ECCV) (2026) 3, 7
53. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Actor and observer: Joint modeling of first and third-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7396–7404 (2018) 4
54. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Charades-ego: A large-scale dataset of paired third and first person videos. arXiv preprint arXiv:1804.09626 (2018) 2
55. Thatipelli, A., Lo, S.Y., Roy-Chowdhury, A.K.: Exocentric to egocentric transfer for action recognition: A short survey. ArXiv preprint arXiv:2410.20621 (2024) 4
56. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models.

ArXiv **abs/2302.13971** (2023), https://api.semanticscholar.org/CorpusID:257219404 4, 10
57. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 8
58. Wang, Q., Zhao, L., Yuan, L., Liu, T., Peng, X.: Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20453–20463 (2023) 4
59. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: Cogvlm: Visual expert for pretrained language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2024) 10, 11
60. Xu, B., Zheng, S., Jin, Q.: Pov: Prompt-oriented view-agnostic learning for egocentric hand-object interaction in the multi-view world. In: Proceedings of the 31st ACM International Conference on Multimedia (MM). pp. 2807–2816 (2023) 4
61. Xu, J., Huang, Y., Hou, J., Chen, G., Zhang, Y., Feng, R., Xie, W.: Retrieval-augmented egocentric video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 3, 4, 14
62. Xu, M., Fan, C., Wang, Y., Ryoo, M.S., Crandall, D.J.: Joint person segmentation and identification in synchronized first- and third-person videos. In: European Conference on Computer Vision (ECCV). pp. 674–693 (2018) 4
63. Xu, M., Gao, M., Gan, Z., Chen, H.Y., Lai, Z., Gang, H., Kang, K., Dehghan, A.: Slowfast-llava: A strong training-free baseline for video large language models. ArXiv preprint arXiv:2407.15841 (2024) 4
64. Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., Kendre, S., Zhang, J., Tseng, S., Lujan-Moreno, G.A., Olson, M.L., Hinck, M., Cobbley, D., Lal, V., Qin, C., Zhang, S., Chen, C.C., Yu, N., Tan, J., Awalgaonkar, T.M., Heinecke, S., Wang, H., Choi, Y., Schmidt, L., Chen, Z., Savarese, S., Niebles, J.C., Xiong, C., Xu, R.: xgen-mm (blip-3): A family of open large multimodal models (2025) 1
65. Xue, Z., Grauman, K.: Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 4
66. Ye, H., Zhang, H., Daxberger, E., Chen, L., Lin, Z., Li, Y., Zhang, B., You, H., Xu, D., Gan, Z., Lu, J., Yang, Y.: Mm-ego: Towards building egocentric multimodal llms for video qa. In: Proceedings of the International Conference on Learning Representations (ICLR) (2025) 8, 9
67. Yu, H., Cai, M., Liu, Y., Lu, F.: What i see is what you see: Joint attention learning for first and third person video co-analysis. In: Proceedings of the 27th ACM International Conference on Multimedia (MM). pp. 1926–1934 (2019) 4
68. Zeng, A., Attarian, M., brian ichter, Choromanski, K.M., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M.S., Sindhwani, V., Lee, J., Vanhoucke, V., Florence, P.: Socratic models: Composing zero-shot multimodal reasoning with language. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=G2Q2Mh3avow 1
69. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE (2023) 1, 10

70. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L., Zhao, D.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025), https://arxiv.org/abs/2501.13106 1, 8, 10, 11
71. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2023) 4, 10
72. Zhang, H., Chu, Q., Liu, M., Shi, H., Wang, Y., Nie, L.: Exo2ego: Exocentric knowledge guided mllm for egocentric video understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026) 12
73. Zhang, R., Gui, L., Sun, Z., Feng, Y., Xu, K., Zhang, Y., Fu, D., Li, C., Hauptmann, A., Bisk, Y., Yang, Y.: Direct preference optimization of video large multimodal models from language model reward. ArXiv arXiv:2404.01258 (2024) 4
74. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data (2024), https://arxiv.org/abs/2410.02713 2, 4