From My View to Yours: Ego-Augmented Learning in Large Vision Language Models for Understanding Exocentric Daily Living Activities

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

[1] University of North Carolina at Charlotte   [2] Salesforce AI Research

https://github.com/dominickrei/EgoExo4ADL

dreilly1@charlotte.edu

Abstract

*Large Vision Language Models (LVLMs) have demonstrated impressive capabilities in video understanding, yet their adoption for Activities of Daily Living (ADL) remains limited by their inability to capture fine-grained interactions and spatial relationships. To address this, we aim to leverage the complementary nature of egocentric views to enhance LVLM's understanding of exocentric ADL videos. Consequently, we propose **ego2exo knowledge distillation** to learn ego-augmented exp representations. While effective, this approach requires paired ego-exo videos, which are impractical to collect at scale. To address this, we propose **Skeleton-guided Synthetic Ego Generation (SK-EGO)**, which leverages human skeleton motion to generate synthetic ego views from exocentric videos. To enhance the ego representation of LVLMs trained on synthetic data, we develop a **domain-agnostic bootstrapped ego2exo** strategy that effectively transfers knowledge from real ego-exo pairs to synthetic ego-exo pairs, while mitigating domain misalignment. We find that the exo representations of our ego-augmented LVLMs successfully learn to extract ego-perspective cues, demonstrated through comprehensive evaluation on six ADL benchmarks and our proposed **Ego-in-Exo PerceptionMCQ** benchmark designed specifically to assess egocentric understanding from exocentric videos.*

1. Introduction

The wide-scale adoption of Large Language Models (LLMs) and availability of large-scale video instruction data has led to the emergence of Large Vision-Language Models (LVLMs) with impressive video understanding capabilities. Learning representations for Activities of Daily Living (ADL) in LVLMs is a particularly promising direction, especially for healthcare applications such as monitoring the elderly, assessing cognitive decline, and assistive robotics.

However, current LVLMs struggle to understand ADL due to two key challenges. First, existing models are primarily trained on large-scale web videos [26, 33] featuring

Figure 1. **LVLM training strategies.** In **Unified ego-exo** a single mapping is learned for both views. In **View-specific ego-exo**, independent representations are learned for each view. **Bootstrapped ego-exo** enables training on synthetic ego-exo datasets.

sports clips [11] and movie scenes [5, 62], which consist mainly of subject-centered frames with prominent motion. This training distribution differs from ADL videos, which contain subtle motions [19] and complex human-object interactions [6]. Second, the practical constraints of collecting ADL data result in datasets predominantly captured from exocentric (exo) cameras. While this view provides a comprehensive view of the scene, it often fails to capture the fine-grained motion and appearance details that are crucial to address the complex challenges of ADL [55].

Inspired by recent works on collecting time-synchronized egocentric (ego) and exo videos [39, 57], we propose learning ego-augmented exo representations in LVLMs to address the challenges of ADL. The ego view naturally captures details of hands and object manipulations, providing fine-grained cues that are often unclear from the exo view. While previous works have explored knowledge transfer from exo to ego representations [40, 69], we investigate the inverse direction and

1

leverage the detailed interaction cues from ego views to enhance exo understanding, as shown in Figure 1. This raises the question: *what strategies are effective for ego-augmented exo representation learning in LVLMs?* We observe that disentangling the representations learned for each viewpoint is more effective than approaches that learn a unified representation space for both views. This disentanglement additionally enables effective cross-view knowledge transfer through a strategy we dub as **ego2exo knowledge distillation (ego2exo)**, which we find to be the most effective way to learn ego-augmented exo representations in LVLMs. This finding is validated through evaluation on various benchmarks designed to measure LVLMs understanding of ADL [56], as well as a novel benchmark, **Ego-in-Exo PerceptionMCQ**, that is specifically designed to measure LVLMs ego understanding from exo videos. Our benchmark is generated through a systematic process leveraging synchronized ego-exo videos from EgoExo4D [28], consisting of 3,881 human-verified multiple-choice questions (MCQs) that probe LVLMs' ability to understand ego cues from exo videos.

LVLMs trained with time-synchronized ego-exo video pairs [28] using ego2exo presents a challenge when training on ADL datasets, where the ego perspective is typically unavailable due to the invasive nature of wearable cameras and the complexity associated with synchronizing cameras across multiple devices. This challenge raises our second question, *how can ego-augmented exo representations be learned when only the exo perspective is available?* Recent methods have attempted to pair unpaired videos by aligning them temporally [73] or using language semantics [70], which works well when ego and exo data being paired shares similar semantics and action distributions, as seen in datasets like ego tennis forehand [73] or between Ego4D [27] and HowTo100M [49]. However, this approach is challenging to apply for ADL, where capturing long, "*boring*" ADL activities is not as common or interesting as recording more engaging activities like *cooking*, which are more likely to be publicly available [18, 35, 79]. Other methods, such as EMBED [23] and Exo2EgoDVC [50], generate synthetic ego perspectives by cropping spatial regions containing human-object interactions from exo videos in HowTo100M. However, these approaches are insufficient for ADL, where hands are not always central to the activity being performed. Instead, the entire human skeleton has proven to be an important modality for understanding ADL understanding [7, 55], as it captures the nuanced body motions characterizing different actions. Consequently, we propose **Skeleton-guided Synthetic Ego Generation (SK-EGO)**, which leverages the motion of human skeleton joints to generate synthetic ego views from exo videos. SK-EGO effectively enables training LVLMs with ego2exo on ADL datasets where only the exo view is available.

SK-EGO enables ego2exo training, but raises the question *can we leverage real ego-exo pairs to learn stronger ego representations for synthetic data?* To address this, we introduce a bootstrapping strategy for LVLMs that transfers representations learned from real ego-exo pairs to enhance training on synthetic pairs. However, this bootstrapping presents a challenge when the real ego-exo pairs come from domains that are misaligned with the target ADL domain, causing ego distribution drift during knowledge transfer (Section 3.3). To address this, we propose **domain-agnostic bootstrapped ego2exo**, consisting of a three-projector architecture and specialized distillation mechanism that preserves learned ego representations while enabling effective knowledge transfer across domains.

To summarize our contributions:
- We introduce the first LVLM that learns ego-augmented exo representations for ADL, enabled through ego2exo knowledge distillation.
- We introduce the EgoPerceptionMCQ benchmark, a human-verified benchmark consisting of 3,881 multiple-choice questions to evaluate LVLMs understanding of ego cues from exo videos.
- We propose SK-EGO, a skeleton-guided method to generating synthetic ego views from exo videos, enabling the training of ego2exo on ADL datasets where collecting ego and exo pairs is impractical.
- We propose domain-agnostic bootstrapped ego2exo, a strategy to effectively transfer knowledge from real ego-exo pairs to synthetic ego-exo pairs while mitigating domain misalignment.

2. Related Work

ADL Representation Learning. While video representation learning has advanced with 3D CNNs [14, 25, 43, 65] and video transformers [4, 8, 24, 41, 47], models optimized on web videos often struggle with complex ADL videos [20, 46, 58, 60, 66]. Human skeleton-based approaches [15, 29, 59, 74] excel in understanding body motion and skeleton action recognition but lack the appearance information needed to model human-object interactions, which is crucial for ADL. To address ADL challenges, several methods combine RGB and pose modalities [1, 21, 34], yet they rely on skeletons at test time, adding computational expense and potential noise in real-world applications. Approaches like π-ViT [55] and VPN++ [22] bypass 3D skeletons at test time through knowledge distillation, transferring information from skeletons to RGB. However, these methods lack the generalized representations of LVLMs and do not leverage the ego view to enhance learning discriminative action representations. In contrast, we are the first to use the ego view to learn action representations for ADL.

Ego-Exo Video Representation Learning. Learning from ego and exo views has been explored in various approaches for video understanding. Prior works can be

categorized [63] into joint-learning and view transfer approaches. Joint learning approaches [50, 61, 70, 71, 73, 75] aim to learn a unified representation space for both views. For example, Actor and Observer [61] trains a dual-stream CNN to contrastively align ego and exo features, while AE2 [73] uses temporal alignment as a contrastive learning objective. In real-world scenarios where only a single view is available for inference, view transfer approaches [3, 40, 52, 54, 67, 69] aim to leverage knowledge from one view to enhance understanding of the other. For example, Ego-Exo [40] uses ego auxiliary tasks to pre-train a 3D-CNN on exo videos. Quattrocchi *et al.* distills knowledge from an exo-trained teacher to an ego student. While these approaches demonstrate the value of cross-view transfer, existing approaches focus on transferring knowledge from exo to ego. In contrast, our work explores the inverse direction of learning ego-augmented exo representations. Furthermore, unlike prior works that focus on traditional video understanding frameworks, we investigate ego-exo representation learning in the context of LVLMs.

Large Vision Language Models for Video Advancements in Large Language Models [10, 16, 64] and large-scale video-text datasets [37, 72, 77, 78] have led to Large Vision Language Models (LVLMs) [32, 36, 42, 48, 76, 77] with impressive video understanding capabilities. While many existing LVLMs contain a mix of ego [27] and exo perspective videos in their training data, the perspectives are not distinguished during training. Our work is the first to investigate how ego-exo video pairs can be used to train LVLMs, validated on exocentric ADL videos.

3. Proposed Method

Preliminary. In this section, we provide an overview of Large Vision Language Models [48] (LVLMs). Consider a video $v \in \mathbb{R}^{T \times H \times W \times 3}$, where T is the number of frames, $H \times W$ is the spatial resolution, and an associated QA pair containing a question q and its corresponding answer a. The video-instruction pairs used to train the LVLM can be denoted as $\mathcal{X} = \{(v_i, q_i, a_i)\}_{i=1}^N$, where the training distribution contains N samples, and $x_i = (v_i, q_i, a_i)$ represents the i'th video-QA pair. A frozen pre-trained visual encoder, CLIP-L/14 [53], is then used to extract visual features from the video v_i, denoted as f_i.

Vicuna [16] is selected as the LLM backbone in the LVLM, with its parameters θ_{LLM} kept frozen. The primary training objective of the LVLM is to achieve vision-language understanding capability through the introduction of visual information into the language model's embedding space. Initially, the visual features f_i do not share a common embedding space with the language model, and a mapping between them must be learned [44]. For this, a learnable feature projector, $\phi(\cdot)$, is used to project the visual features f_i into the embedding space of the language model.

The projected visual features $\phi(f_i)$ and query q_i are then input to the language model following the template:

> USER: <q_i> <$\phi(f_i)$> Assistant:

During training, the language model iterates over samples in the video-QA pairs, $\mathcal{X}$, and processes each video-question pair to generate next token predictions. The LVLM is trained using an auto-regressive loss as

$$\mathcal{L}_{\text{LLM}} = -\sum_{t=1}^{T} \log Pr(x_t \mid x_{<t}; \theta_{LLM}) \qquad (1)$$

where T is the length of the input sequence and $Pr(x_t \mid x_{<t}; \theta)$ is the probability of the token x_t given the preceding tokens $x_{<t}$ (all tokens before x_t).

Overview. In the typical LVLM training paradigm, all videos are processed identically regardless of their view (ego or exo). This results in LVLMs that fail to leverage the complementary visual cues available between ego and exo views. In contrast, we take advantage of these complementary cues to learn ego-augmented exo representations in LVLMs, enabling them to infer ego cues from exo videos at inference when only exo videos are available. Consequently, in this section we (1) propose various strategies to learn ego-augmented exo representations in LVLMs, (2) present SK-EGO, a skeleton-guided cropping strategy for generating synthetic ego views in exo-only datasets, and (3) propose a mechanism to transfer knowledge from real to synthetic ego-exo pairs for training LVLMs in ADL.

3.1. Strategies for Learning Ego-augmented Exo Representations in LVLMs

Here, we assume the availability of time synchronized ego-exo videos (real ego-exo) for training, resulting in the video-instruction pairs, $\mathcal{X}^{\text{egoexo}} = \{x_i^{\text{egoexo}} = (v_i^{\text{ego}}, v_i^{\text{exo}}, q_i, a_i)\}_{i=1}^N$, where v_i^{ego} and v_i^{exo} correspond to synced videos captured from the ego and exo views. Let f_i^{ego} and f_i^{exo} denote the corresponding visual features extracted from a frozen pre-trained visual encoder. This setting with real ego-exo video pairs provides an ideal test bed for evaluating different strategies. The strategies we explore in this section are illustrated in Figure 2.

Unified ego-exo representation. This strategy adopts the vanilla LVLM architecture consisting of a single feature projector, $\phi^v(\cdot)$ where v indicates the view(s) used during training, that learns a mapping from visual features to the embedding space of the language model. In this way, a unified representation space is learned for both views.

Disentangled ego-exo representation. This strategy deploys two distinct feature projectors, one ego view projector ϕ^{ego} and one exo view projector ϕ^{exo}. After obtaining visual features for each view, f_i^{ego} and f_i^{exo}, they are passed to their respective projectors and input jointly to the language model along with the query q_i using the following template:

> USER: <q_i> <$\phi^{\text{ego}}(f_i^{\text{ego}})$> <$\phi^{\text{exo}}(f_i^{\text{exo}})$> Assistant:

Figure 2. **LVLM training strategies explored in this work.** **(a)** Unified ego-exo representation shares a feature projector for ego and exo views. **(b)** Disentangled ego-exo representation learns independent feature projectors for each view through joint ego-exo training, maintaining view-specific representations. **(c)** Ego2exo distillation bi-directionally transfers knowledge between ego/exo feature projectors.

Figure 3. **SK-EGO: Skeleton-guided ego view generation.** SK-EGO computes motion magnitudes across skeleton joints to identify regions of significant activity in exo videos. The joints with highest temporal motion guide the cropping of ego-like views.

Unlike unified ego-exo representations, this strategy enables learning independent representations for each view.

Knowledge distillation strategies. Knowledge distillation (KD) enables the transfer of knowledge from one neural network to another. As we aim to learn ego-augmented exo representations, distillation serves as a natural strategy to learn egocentric cues in exocentric representations. We propose to introduce distillation into the disentangled ego-exo representation learning strategy, denoted as **ego2exo** and illustrated in Figure 2. Specifically, the projector outputs of the ego feature projector, $\phi^{\text{ego}}(f_i^{\text{ego}})$, and the exo feature projector, $\phi^{\text{exo}}(f_i^{\text{exo}})$, are bi-directionally distilled to one another during training. Both projectors remain trainable throughout this process, facilitating mutual knowledge transfer between the viewpoints. The total loss for the ego2exo LVLM is a convex optimization of a distillation loss and $\mathcal{L}_{LLM}$, defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dist}}(\phi^{\text{ego}}(f_i^{\text{ego}}), \phi^{\text{exo}}(f_i^{\text{exo}})) + \mathcal{L}_{\text{LLM}} \quad (2)$$

$$\mathcal{L}_{\text{dist}}(A, B) = \|\Sigma_1^\tau A - \Sigma_1^\tau B\|^2 \quad (3)$$

where τ corresponds to the number of visual tokens. To validate the effectiveness of these strategies, we develop the Ego-in-Exo Perception MCQ benchmark specifically designed to measure egocentric understanding in exocentric videos. This benchmark is presented in Section 4.

3.2. Skeleton-guided Synthetic Ego Generation for ADL (SK-EGO)

Most existing ADL datasets capture only exo views, preventing the adoption of representation learning strategies re-

quiring both ego and exo views. To address this, we propose **SK-EGO**: a skeleton-guided cropping strategy to generate a *synthetic egocentric* view of exocentric videos, as shown in Figure 3. We argue that 2D skeletons effectively characterize the motion present in ADL and can be easily obtained using off-the-shelf pose estimators [13].

The generated synthetic ego, while not a true ego view, aims to approximate the key cues emphasized in the ego view, such as fine-grained motions and HOIs. Existing cropping-based approaches to synthetic ego generation focus only on capturing HOIs [23], but in ADL scenarios, this is insufficient as many activities involve broader body movements beyond HOIs. In contrast to HOI-guided approaches, SK-EGO extracts crops from spatial regions containing joints with the highest motion, shifting focus from hands to other active body parts when HOIs are absent.

SK-EGO is guided by the motion of the human skeleton joints over time. Specifically, let $\mathbf{S}_i \in \mathbb{R}^{T \times J \times 2}$ represent the 2D skeleton sequence of the video v_i^{exo}, containing T frames and the 2D spatial coordinates of J human joints. We first compute motion magnitudes of each human joint across the video as

$$\mathcal{M}_i = \frac{1}{T} \sum_{t=1}^{T-1} \|\mathbf{S}_i^{t+1} - \mathbf{S}_i^t\|_2 \quad (4)$$

where $\mathcal{M}_i \in \mathbb{R}^J$ represents the motion magnitudes of the joints in the video, and $\mathbf{S}_i^t$ are the 2D skeleton joints at frame t. Prior to computing $\mathcal{M}_i$, skeletons are centered and normalized with respect to the first frame. Then, the coordinates of the Top-K joints with the largest motion magnitude are used to extract a spatial crop from the exo video, generating a synthetic ego video. Thus, the synthetic ego video, $v_i^{\overline{\text{ego}}}$, can be computed from $\mathcal{M}_i$ and v_i^{exo} as follows

$$\mathcal{M}_i^k = \text{Top-}K(\mathcal{M}_i); \qquad v_i^{\overline{\text{ego}}} = \text{Crop}(v_i^{\text{exo}}, \mathcal{M}_i^k) \quad (5)$$

where $\mathcal{M}_i^k \in \mathbb{R}^k$ are the k joints in video i with the largest motion over time and Crop(.) returns the minimum spanning bounding box that encapsulates all the joints of $\mathcal{M}_i^k$ across all frames in v_i^{exo}.

In summary, given an ADL dataset consisting exclusively of exocentric videos, SK-EGO returns synthetic ego-exo, $\mathcal{X}^{\overline{\text{ego}}\text{exo}} = \{(v_i^{\overline{\text{ego}}}, v_i^{\text{exo}}, q_i, a_i)\}_{i=1}^N$, where v_i^{exo} is the real exo video and $v_i^{\overline{\text{ego}}}$ is the synthetic ego video. The generated synthetic ego-exo pairs enable the training of rep-

Figure 4. **Domain-agnostic bootstrapped ego2exo**. Knowledge transfer is performed between a bootstrapping projector, and the fusion of ego and exo projectors. Ego projector is initialized from ego2exo trained on real ego-exo pairs.

resentation learning strategies discussed in Section 3.1 in scenarios when only exo videos are available for training. In the next section, we introduce an effective strategy for transferring knowledge from real ego-exo pairs to synthetic ego-exo pairs to enhance LVLM's performance for domain-specific ADL downstream tasks.

3.3. Domain-Agnostic Bootstrapping Synthetic Ego-Exo Training with Real Ego-Exo Data

A naive approach to transferring knowledge from real to synthetic ego-exo pairs involves either joint training or progressive strategy – first bootstrapping with real pairs, followed by learning representations for synthetic ego-exo pairs. However, real ego-exo pairs may originate from instructional videos or robotics domains, which may misalign with the target ADL domain. This domain shift can hinder the LVLM's ability to learn domain-specific representations effectively. To address this challenge, we propose a domain-agnostic bootstrapping strategy that effectively transfers ego-augmented representations from real to synthetic ego-exo pairs.

Bootstrapping for LVLM Training. We introduce a two-stage progressive training mechanism to bootstrap the LVLM with real ego-exo video instructions before training on synthetic ego-exo video instructions.

Stage 1: Learning Ego Representations from Real Ego-Exo Pairs. We first train an ego2exo LVLM for one epoch on the EgoExo4D dataset, enabling the model to learn a rich ego representation from real ego-exo pairs.

Stage 2: Bootstrapping Ego2Exo. To transfer the learned ego representations to synthetic ego-exo data, we introduce an additional bootstrapping projector for the exo view, denoted ϕ^{boot}, to learn ADL-specific representations and distill knowledge from the ego representations learned in Stage 1. The *ego* projector is initialized with weights learned from Stage 1 training, whereas the exo and bootstrapping projectors are initialized from LLaVA [45], since only the ego view is synthetic in Stage 2.

Domain-Agnostic Bootstrapped Ego2Exo Distillation. During Stage 2, applying standard knowledge distillation (i.e., ego2exo) fails to effectively transfer representations learned from real ego-exo pairs due to ego distribution drift, caused by the direct distillation between ego and exo projectors. To mitigate this, we propose domain-agnostic **boot-**

Figure 5. **The strategy for generating Ego-in-Exo Perception-MCQs from paired ego-exo videos.** The atomic action descriptions and scene object list are fed to a Large Language Model (GPT-4o), which generates a single question for each of the four question categories. MCQs are manually verified by humans before being included in the final version of the benchmark.

strapped ego2exo distillation, as illustrated in Figure 4. Our method preserves the strong ego distribution learned from real ego-exo pairs by enforcing that the fusion of ego and exo representations remains closer to the bootstrapping exo projector's representations. This fusion acts as a balancing factor, countering distribution drift introduced by the distillation loss. The training loss for bootstrapped ego2exo is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dist}}(\phi^{\text{boot}}(f_i^{\text{exo}}), (\phi^{\text{ego}}(f_i^{\text{ego}}) + \phi^{\text{exo}}(f_i^{\text{exo}}))) + \mathcal{L}_{\text{LLM}} \quad (6)$$

During training, both the output from the bootstrapping projector and the fused representation are fed to the language model following the template:

```
USER: <q_i> <φ^boot(f_i^exo)> <φ^ego(f_i^ego) + φ^exo(f_i^exo)>
                Assistant:
```

4. Ego-in-Exo PerceptionMCQ

To benchmark the ability of LVLMs to understand ego cues from exo videos, we introduce Ego-in-Exo Perception-MCQ, a large-scale multiple-choice question benchmark derived from EgoExo4D [28]. While EgoExo4D contains long video takes of skilled human activities, we utilize the keystep clips - short temporal segments from long videos that capture specific procedural actions. Each keystep clip is annotated with a descriptive label and synchronized across an ego and multiple exo views. Further, we only consider the subset of "*cooking*" keysteps in the dataset. This subset provides 2319 unique keystep videos along with 10K timestamped atomic action descriptions detailing the fine-grained activities within each keystep. The key idea of our benchmark is to generate MCQs from the *ego-view* and evaluate them on the paired *exo-view*. Figure 5 illustrates the creation of Ego-in-exo PerceptionMCQ.

Our benchmark comprises four question categories (Action Understanding, Task-relevant Region, Human Object Interactions, and Hand Identification) designed to evaluate understanding of various ego cues. We enhance the annotations by extracting scene objects using an image captioner [30] with a sliding window over the ego viewpoint, and feed the scene objects along with corresponding EgoExo4D annotations to a large language model to generate MCQs. Questions are generated using GPT-4o [51] with category-specific prompts, after which we conduct rigorous human verification to ensure quality. Additional details on benchmark construction are available in the supplementary.

5. Experiments

In this section, we first present the evaluation settings and then provide a thorough analysis of learning ego-augmented exo representation in LVLMs.

5.1. Evaluation Settings

Datasets. EgoExo4D [28] is a large-scale multi-view dataset containing over 1,200 hours of time-synchronized ego and exo videos. As the dataset lacks instruction tuning data required to train LVLMs, we generate ego-exo Video-QA pairs from keystep activity clips and atomic action descriptions provided as annotations in EgoExo4D, only considering the highest quality exo videos as indicated by the annotations. More details are provided in the supplementary materials. We will release these 50K instruction tuning pairs to promote future research.

ADL-X [56] is an instruction tuning dataset designed for ADL-focused LLVMs, containing over 100k instruction tuning pairs. The dataset is created through a weakly supervised data curation framework that provides video QA pairs for temporally stitched videos from NTU120 [46]. We denote the ADL-X dataset with our proposed SK-EGO generated synthetic ego views as ADL-X-EgoExo.

Downstream Tasks. We evaluate our methods across 10 benchmarks designed to measure LVLM's ability to understand (1) ADL and (2) egocentric understanding from exo videos. ADL Multiple Choice Questions (ADL-MCQ) [56] consists of four benchmarks to assess the question answering ability of LVLMs on ADL questions, and ADL Video Description (ADL-VD) [56] contains two benchmarks to measure description capability. Raw accuracy is reported for ADL-MCQ (consisting of Temporal Completion (TC) and Action Recognition (AR) question types), and Video-ChatGPT description metrics [48] are reported for ADL-VD. To measure egocentric understanding of LVLMs on exo videos, we report accuracy on the four categories of our proposed Ego-in-exo PerceptionMCQ benchmark.

Implementation Details. In all of our experiments, Vicuna 1.1 [16] is used as the backbone LLM and CLIP-L/14 [53] is used as the visual encoder. Following [48],

we perform spatio-temporal pooling on the encoded visual features. Regardless of view, this pooling results in a total of $\tau = 356$ visual tokens per video. While training, both the visual encoder and LLM are kept frozen and only the feature projectors are trainable. The LLM and visual encoder are initialized with parameters from LLaVA [44]. All experiments are trained on 8 A6000 48GB GPUs for 3 epochs with a total batch size of 32 and a learning rate of $2e^{-5}$. When applying SK-EGO to ADL-X, we set $K = 6$ for selecting joints with the largest motion.

5.2. Discussion and Analysis

Which strategy is effective for Ego-augmented Exo Representation Learning? Table 1 presents the results of (1) Unified, (2) Disentangled, and (3) Knowledge Distillation-based ego-exo representation learning strategies. The methods are trained on real ego-exo pairs from EgoExo4D (EE4D) and synthetic ego-exo pairs from ADL-X-EgoExo (ADLX-EE), and are evaluated on ADL-MCQ. We find that the disentangled ego-exo representations consistently outperform unified representations (36.5% vs 35.5% on MCQ Avg when trained on EgoExo4D), emphasizing the importance of learning dissociated representations through view-specific feature projectors. As for distillation strategies, we evaluate the ego2exo strategy along with an offline variant, in which a pre-trained and frozen LVLM trained on ego videos is used as a teacher for an exo-trained LVLM student. We observe that ego-exo representations learned through knowledge distillation outperform disentangled representation learning strategies, with ego2exo achieving an average accuracy on ADL MCQ of 37.4% vs 36.5% when trained on EE4D, and 43.4% vs 42.5% when trained on ADLX-EE.

Are exo representations learning ego cues? To answer this question, we evaluate ego-exo trained LVLMs on our proposed Ego-in-exo PerceptionMCQ benchmark and present the results in Table 2. As an upper bound, we present the results of the unified ego-exo LVLM trained on ego videos from EgoExo4D, and evaluated on the ego view of the benchmark, all other methods are evaluated on the exo view. Consistent with our findings in Table 1, ego2exo demonstrates superior performance across all question categories, achieving 40.3% average accuracy on the benchmark, a +4.4% improvement over the disentangled representation LVLM. This improvement is particularly significant for the HOI category (37.8% vs 33.5%), suggesting enhanced understanding of ego cues in the ego2exo LVLM.

Which strategy is most effective for generating synthetic ego views? In Table 3, we evaluate different strategies for generating synthetic ego views, comparing them to our proposed SK-EGO approach for exo-only datasets. First, we explore diffusion-based approaches using OpenAI's DALLE-3 [9]. We use action labels from the stitched ADL-X videos and prompt the DALLE model to generate first-person views of the actions. Additionally, we exam-

Table 1. **Comparison of LVLM training strategies.** We report the ADL-MCQ accuracy of representation learning strategies trained on the EgoExo4D and ADL-X-EgoExo datasets. Single-view LVLM denotes a vanilla LVLM trained only on exo-view videos. Offline ego2exo denotes ego2exo training when the ego feature projector is not trainable. The highest average accuracy on ADL-MCQ is bolded.

Method	Train views		Trained on EgoExo4D (EE4D)					Trained on ADL-X-EgoExo (ADLX-EE)				
	Ego	Exo	Charades AR	SH AR	LEMMA TC	TSU TC	MCQ Avg	Charades AR	SH AR	LEMMA TC	TSU TC	MCQ Avg
Unified ego-exo	✓	✓	45.2	30.5	32.6	33.8	35.5	51.7	43.7	28.3	30.2	38.5
Single-view LVLM	✗	✓	42.8	30.8	33.7	36.3	35.9	51.0	44.5	28.6	29.5	38.4
Disentangled ego-exo	✓	✓	43.5	31.2	36.3	35.1	36.5	54.7	44.3	35.0	35.9	42.5
Ego2exo (Offline)	✓	✓	46.5	33.6	33.0	34.6	36.9	52.3	45.6	34.1	36.5	42.1
Ego2exo	✓	✓	45.3	33.6	36.6	34.2	**37.4**	54.1	47.6	33.7	38.2	**43.4**

Table 2. **Evaluation on Ego-to-exo PerceptionMCQ.** We validate the explored ego-augmented exo representation learning strategies on our proposed benchmark. Ego2exo demonstrates the highest ego-in-exo understanding ability, consistent with Table 1.

Method	Training Views	Action	Task-R	HOI	Hand	Avg
Evaluate on ego view (upper bound)						
Single-view LVLM	Ego	31.0	50.5	44.3	51.0	44.2
Evaluate on exo view						
ADL-X-ChatGPT [56]	Exo	15.2	31.5	32.8	52.1	32.9
LLAVIDAL [56]	Exo	15.2	33.3	31.0	53.8	33.3
Disentangled ego-exo	Ego-exo	15.4	43.2	33.5	51.4	35.9
Ego2exo	Ego-exo	**17.5**	**49.0**	**37.8**	**57.0**	**40.3**

Table 3. **Synthetic ego generation strategy.** We evaluate cropping and diffusion-based strategies for synthetic ego generation on ego2exo. SK-EGO performs best for ADL understanding tasks.

Method	Synthetic ego generation strategy	Charades AR	SH AR	TSU TC	MCQ Avg	Charades Desc. Avg
Ego2exo	DALLE-3	52.1	45.4	37.7	45.1	45.2
	DALLE-3 (+ scene desc.)	54.5	45.5	38.4	46.1	46.4
	EMBED [23]	52.7	45.8	38.0	45.5	47.9
	SK-EGO	54.1	47.6	38.2	**46.6**	**48.2**

Table 4. **Bootstrapping LVLMs with real ego-exo pairs.** EgoExo4D (EE4D) consists of real ego-exo pairs, ADL-X-EgoExo (ADLX-EE) consists of synthetic ego-exo pairs. EE4D (20%) indicates (20%) of the ego-exo instruction pairs were used.

Method	Training Data	Charades AR	SH AR	TSU TC	MCQ Avg	Charades Desc. Avg
Unified ego-exo	EE4D + ADLX-EE	48.2	46.4	36.4	43.7	42.9
Ego2exo	EE4D	50.6	35.6	39.4	41.9	41.8
	ADLX-EE	54.1	47.6	38.2	46.6	48.2
	EE4D + ADLX-EE	52.8	47.6	40.3	46.9	46.6
	EE4D (20%) + ADLX-EE	50.7	47.4	40.7	46.2	45.4
+ bootstraping	Epoch 1: EE4D Epoch 2-3: ADLX-EE	53.6	47.4	39.9	47.0	46.4
+ Domain-Agnostic bootstraping	Epoch 1: EE4D Epoch 2-3:ADLX-EE	**55.0**	**48.1**	**49.9**	**48.0**	**48.7**

ine a variant that incorporates scene context into the generation prompt. We also evaluate EMBED [23], which adopts an HOI-guided spatial cropping strategy. We find that diffusion-based methods can perform well, but they are limited by quality, cost, and inconsistent generation quality. Compared to EMBED, our proposed SK-EGO approach achieves the highest performance (46.6% vs. 45.5% on ADL-MCQ Avg). This is attributed to EMBED's exclusive focus on HOIs, which fails to capture the full spectrum of activities present in ADL.

How to best bootstrap training on synthetic ego-exo with real ego-exo data? We explore approaches for bootstrapping LVLMs trained on synthetic ego-exo pairs with real ego-exo pairs in Table 4. Consistent with our previous find-

ing, ego2exo trained on a naive combination of real and synthetic ego-exo pairs outperforms unified ego-exo representations (43.7% vs 46.9% on MCQ Avg). We also find that bootstrapping applied to ego2exo performs comparably to ego2exo trained on the combination of real and synthetic ego-exo pairs (47.0% vs 46.9% on MCQ Avg and 46.3% vs 46.6% on Charades Desc). The poor performance of bootstrapping on ego2exo is attributed to the ego distribution drift discussed in Section 3.3. Notably, we observe that our proposed domain agnostic bootstrapping mitigates this issue, achieving significant performance improvements over ego2exo + bootstrapping (48.0% vs 47.0% on MCQ Avg and 48.7% vs 46.4% on Charades Desc).

6. Comparison to the state-of-the-art

Table 5 presents results comparing our bootstrapped ego2exo LVLM training against existing LVLMs and the state-of-the-art on ADL understanding, including two-stage approaches combining image captioning with LLMs. Image-language models such as CogVLM, even when paired with strong language models (GPT-4), achieve limited performance (53.3% on TSU Description) compared to our ego-augmented LVLMs (73.9%). Existing LVLMs trained on web videos struggle to understand ADL when compared to bootstrapped ego2exo LVLM training (48.1% on SH-AR vs 39.6% with Video-ChatGPT), highlighting the gap between web and ADL videos. Bootstrapped ego2exo training significantly outperforms the most representative baseline ADL-X-ChatGPT, achieving 48.1% vs 44.5% on SH-AR and 34.5% vs 28.6% on LEMMA-AF. Notably, our method outperforms LLAVIDAL, despite using only training the LVLM with RGB inputs compared to LLAVIDAL's use of an additional depth modality, demonstrating the effectiveness of ego-augmented exo representation learning for understanding ADL with LVLMs.

7. Qualitative Results

Ego-Exo Feature Distances. Figure 6 (left) shows the distances between feature representations of ego-exo video pairs. We compare two approaches: (1) our ego2exo LVLM, and (2) a unified representation LVLM. For ego2exo, we encode both views using the CLIP video encoder and pass them through their respective ego/exo fea-

Table 5. **Comparison with state-of-the-art methods on ADL understanding.** Image-language paired with web-trained models serve as general vision-language baselines. ADL-X-ChatGPT and LLAVIDAL represent domain-specific approaches trained on ADL instruction data. Modality QA pairs indicates the total number of instruction tuning pairs multiplied by the number of modalities present in each pair.

Method	# Training Modality QA Pairs	ADL MCQ					Charades Description						Toyota Smarthome Description					
		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 [68] + GPT [10]	N/A	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 [68] + Llama [64]	N/A	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 [38] + GPT [10]	N/A	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
Web video trained LVLMs																		
Video-ChatGPT [48]	100K	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
Video-LLaMA [76]	2.6M	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
Video-LLaVA [42]	1.2M	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 [32]	**3M**	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 trained LVLMs																		
LLAVIDAL [56]	300K	**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
ADL-X-ChatGPT [56]	100K	51.0	44.5	28.6	29.5	38.4	40.4	50.6	49.8	30.6	40.2	42.4	62.4	79.4	70.8	51.2	60.4	64.8
Bootstrapped ego2exo (Ours)	150K	55.0	48.1	34.5	**40.9**	**44.6**	45.2	64.4	56.8	34.3	42.8	**48.7**	68.6	91.2	84.8	55.4	69.4	**73.9**

Figure 6. (**Left**) Feature distances between ego and exo view videos. The ego video is denoted with a blue border, the unified ego-exo LVLM with a orange border, and the ego2exo LVLM with a green border. Distance is computed between the ego and exo video representations of respective methods. (**Right**) Average motion magnitudes of the 25 skeleton joints in ADL-X.

Figure 7. **Qualitative LVLM outputs.** Results are shown for Video-ChatGPT, ADL-X-ChatGPT, and Bootstrapped ego2exo on the ADL-MCQ (Smarthome Action Recognition) and ADL-Video Descriptions (Charades) benchmarks.

ture projectors before computing Euclidean distances. For the unified representation LVLM, both videos are encoded and passed through the same projector before distance calculation. Our ego2exo LVLM achieves consistently smaller distances between corresponding ego-exo pairs than the unified representation LVLM, demonstrating implicit ego representation learning within exo representations through our ego2exo distillation method.

SK-EGO Joint Selections. Figure 6 (right) visualizes the average motion magnitudes of the joints in ADL-X, as computed by SK-EGO. For each video, the per-joint motion magnitudes are computed and then averaged across all videos in the dataset to obtain the magnitude for each joint. Darker colors indicate higher motion magnitudes and are thus more likely to be selected by SK-EGO. This analysis demonstrates that SK-EGO is better suited for ADL, as HOI-guided synthetic ego generation strategies [23] fail to capture the full spectrum of activities in ADLs.

Example LVLM Answers. Figure 7 compares the qualitative results of three LVLMs: Video-ChatGPT [48], trained on web videos; ADL-X-ChatGPT [56], trained on ADL videos; and our bootstrapped ego2exo LVLM. The left side of the Figure compares the answers of the three models on an example from ADL-MCQ (SH-AR). The right side

demonstrates the effectiveness of our method in generating detailed responses to open-ended descriptive questions.

8. Conclusion

In this paper, we address the unexplored area of training LVLMs for understanding exocentric ADL using egocentric views. We explored various strategies for integrating ego-view cues into exo representations, finding ego2exo knowledge distillation to be most effective. This was validated on ADL-X benchmarks, as well as our proposed **Ego-in-Exo PerceptionMCQ** benchmark designed to assess LVLM's understanding of ego cues from exo videos. To overcome the practical challenge of limited paired ego-exo data in real-world ADL scenarios, we developed **Skeleton-guided Synthetic Ego Generation (SK-EGO)**, which generates synthetic ego views from exo videos, guided by human skeleton motion. To learn stronger ego representations when training on these synthetic ego-exo pairs, we propose **domain-agnostic bootstrapped ego2exo**, a novel strategy that effectively transfers knowledge from real ego-exo pairs to synthetic pairs while mitigating domain misalignment. This is the first attempt towards learning ego-augmented representations for ADL, demonstrating the potential of ego-exo perspectives for learning discriminative ADL representations and warranting future explorations in this area.

Acknowledgments

References

[1] Dasom Ahn, Sangwon Kim, Hyun Wook Hong, and ByoungChul Ko. Star-transformer: A spatio-temporal cross attention transformer for human action recognition. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3319–3328, 2023. 2

[2] Meta AI. The llama 3 herd of models, 2024. 14

[3] Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. In *Computer Vision and Image Understanding*, pages 61–68, 2018. 3

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021. 2

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 1

[6] Fabien Baradel, Christian Wolf, and Julien Mille. Human activity recognition with pose-driven attention to rgb. In *The British Machine Vision Conference (BMVC)*, 2018. 1

[7] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W. Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 2

[9] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. In *OpenAI Blog*, 2023. 6

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3, 8

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 1

[12] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv Preprint arXiv:2410.10818*, 2024. 13

[13] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 4

[14] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4724–4733. IEEE, 2017. 2

[15] Hyung-Gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20154–20164, 2022. 2

[16] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3, 6

[17] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024. 13

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[19] Srijan Das, Arpit Chaudhary, Francois Bremond, and Monique Thonnat. Where to focus on for human action recognition? In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 71–80, 2019. 1

[20] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Int. Conf. Comput. Vis.*, 2019. 2

[21] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *European Conference on Computer Vision*, pages 72–90. Springer, 2020. 2

[22] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2

[23] Zi-Yi Dou, Xitong Yang, Tushar Nagarajan, Huiyu Wang, Jing Huang, Nanyun Peng, Kris Kitani, and Fu-Jen Chu. Un-

locking exocentric video-language data for egocentric video representation learning, 2024. 2, 4, 7, 8

[24] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 2

[25] Christoph Feichtenhofer. X3D: expanding architectures for efficient video recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 200–210. Computer Vision Foundation / IEEE, 2020. 2

[26] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5843–5851. IEEE Computer Society, 2017. 1

[27] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 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)*, pages 18995–19012, 2022. 2, 3

[28] Kristen Grauman et al. 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 (CVPR)*, 2024. 2, 5, 6

[29] Ryo Hachiuma, Fumiaki Sato, and Taiki Sekii. Unified keypoint-based action recognition framework via structured keypoint pooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22962–22971, 2023. 2

[30] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6, 13

[31] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models, 2025. 13

[32] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 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. 3, 8

[33] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[34] Sangwon Kim, Dasom Ahn, and Byoungchul Ko. Cross-modal learning with 3d deformable attention for action recognition. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[35] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 2

[36] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *ArXiv preprint arXiv:2408.03326*, 2024. 3

[37] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 3

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 8

[39] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6943–6953, 2021. 1

[40] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. 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)*, pages 10995–11005, 2021. 1, 3

[41] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 2

[42] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. 3, 8

[43] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3, 6

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 5

[46] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 6

[47] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, 2021. 2

[48] Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 3, 6, 8

[49] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 2

[50] Takehiko Ohkawa, Takuma Yagi, Taichi Nishimura, Ryosuke Furuta, Atsushi Hashimoto, Yoshitaka Ushiku, and Yoichi Sato. 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. 2, 3

[51] OpenAI. Gpt-4o system card, 2024. 6, 13

[52] Camillo Quattrocchi, Antonino Furnari, Daniele Di Mauro, Mario Valerio Giuffrida, and Giovanni Maria Farinella. 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. 3

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 3, 6

[54] Arushi Rai, Kyle Buettner, and Adriana Kovashka. 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*, pages 1714–1723, 2024. 3

[55] Dominick Reilly and Srijan Das. Just add π! pose induced video transformers for understanding activities of daily living. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[56] Dominick Reilly, Rajatsubhra Chakraborty, Arkaprava Sinha, Manish Kumar Govind, Pu Wang, Francois Bremond, Le Xue, and Srijan Das. Llavidal: A large language vision model for daily activities of living. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 6, 7, 8

[57] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1

[58] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2

[59] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020. 2

[60] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *European Conference on Computer Vision(ECCV)*, 2016. 2

[61] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7396–7404, 2018. 3

[62] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640, 2015. 1

[63] Anirudh Thatipelli, Shao-Yuan Lo, and Amit K. Roy-Chowdhury. Exocentric to egocentric transfer for action recognition: A short survey. *ArXiv preprint arXiv:2410.20621*, 2024. 3

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. 3, 8

[65] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society. 2

[66] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[67] Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20453–20463, 2023. 3

[68] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 8

[69] Boshen Xu, Sipeng Zheng, and Qin Jin. 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)*, pages 2807–2816, 2023. 1, 3

[70] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[71] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S. Ryoo, and David J. Crandall. Joint person segmentation and identification in synchronized first- and third-person videos. In *European Conference on Computer Vision (ECCV)*, pages 674–693, 2018. 3

[72] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *ArXiv preprint arXiv:2407.15841*, 2024. 3

[73] Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3

[74] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 2

[75] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. 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)*, pages 1926–1934, 2019. 3

[76] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 3, 8

[77] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward. *ArXiv arXiv:2404.01258*, 2024. 3

[78] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *ArXiv arXiv:2410.02713*, 2024. 3

[79] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

From My View to Yours: Ego-Augmented Learning in Large Vision Language Models for Understanding Exocentric Daily Living Activities

Supplementary Material

9. Appendix

Figure 8. Ego-in-Exo Perception MCQ Statistics. We present the MCQ count of each category, along with the per-category unique word count and average MCQ length.

9.1. Additional Ego-in-Exo PerceptionMCQ Details

Detailed statistics of Ego-in-Exo PerceptionMCQ can be found in Figure 8. In the following, we provide additional details on the construction of our Ego-in-Exo Perception-MCQ benchmark.

Scene object parsing. Atomic action descriptions provide dense details about actions and HOIs, but they lack scene content. Extracting scene objects enables the generation of more diverse question types, and facilitates the creation of more challenging negative distractors. To obtain scene objects for a given keystep clip, we use an CogAgent [30] with a sliding window over the ego viewpoint, generating image captions at 5fps. The prompt we use for CogAgent is "Describe the scene and what objects are visible in the scene". We then use an LLM to parse the captions

into a list of scene objects, which we use as input to GPT-4o for generating MCQs.

MCQ Generation. Our benchmark comprises *four* distinct categories of multiple-choice questions, manually designed to evaluate understanding of different ego cues. For each keystep clip, we leverage GPT-4o [51] with category-specific prompts to generate a single question per category. The four categories are defined as: **Action Understanding (Action)**, which assesses comprehension of the actions being performed; **Task-relevant Region (Task-R)**, which evaluates understanding of spatial areas where the primary keystep action is being performed; **Human Object Interactions (HOI)**, which measures understanding of human and object interactions; and **Hand Identification (Hand)**, which evaluates ability to distinguish the specific hand used to perform an action. Depending on the category we provide different annotations to GPT-4o. These four category-specific prompts will be open-sourced along with the Ego-in-Exo PerceptionMCQ benchmark.

Human quality verification Previous research [12, 17, 31] has identified common issues with LLM-generated MCQs, including hallucinated questions and easy negative answers. Our benchmark presents an additional unique challenge: as questions are generated using annotations derived from the ego view, they may not always be answerable from the exo perspective due to occlusions or camera placement. To address these concerns, we first filter out MCQs that the LLM could not confidently generate, resulting in 5,689 MCQs. We then employ four human annotators to manually verify these MCQs on a scale of 1-3, retaining only those with a score of 3, resulting in a total of 3,881 MCQs used in our benchmark. Figure 10 provides a glimpse of the Human Annotation tool.

Number of Joints (k)	Charades AR	SH AR	TSU TC	Charades Desc.
4	52.0	50.3	27.6	47.9
6 (default)	52.9	50.7	28.1	48.6
8	51.7	50.7	30.6	48.3

Table 6. **Ablation on SK-EGO number of joints.**

9.2. Analysis of SK-EGO: Top-k Joint Selection

In this section, we discuss the top-k joint selection of SK-EGO and explore the optimal choice of k.

In Table 6, we ablate the number of skeleton joints selected by SK-EGO for cropping. While the model performance remains relatively stable across different values of k, we find that $k = 6$ performs best on average. This suggests

Figure 9. **Generating instruction data for tuning LVLMs on EgoExo4D.** We use the keystep segments and corresponding atomic action narrations from EgoExo4D to generate instruction pairs for training LVLMs.

Figure 10. **Verification interface.** Human verifiers were asked to rate generated video-MCQ pairs on a scale of 1-3.

a trade-off: too few joints may miss crucial interaction regions, while too many lead to overly large crops that dilute the ego-like perspective.

9.3. Synthetic Ego Generation using DALL-E

We generate synthetic ego views from exo videos using OpenAI's DALLE-3 diffusion model. For each action in a temporally stitched ADL-X video, we create a corresponding ego image, then stitch these images into a synthetic ego video that pairs with the original exo video. We evaluate two prompting strategies to DALLE: (1) using only the action name, or (2) combining the action name with scene descriptions obtained from an image captioning model. The resulting synthetic ego-exo pairs are processed through the LVLM following the same LVLM pipeline as real ego/exo videos. Figure 11 presents visualizations of the diffusion

Figure 11. **Qualitative visualization of synthetic ego views.** We generate synthetic ego views using DALLE-3 with/without scene descriptions, and our proposed SK-EGO.

generated synthetic ego views.

9.4. EgoExo4D Data Generation Pipeline

While EgoExo4D provides synchronized ego-exo video pairs, it lacks the instruction-tuning data required to train our ego-augmented LVLMs. To address this, we develop a pipeline to automatically generate high-quality instruction tuning data from EgoExo4D. Our pipeline is illustrated in Figure 9 and leverages a large language model (Llama 3.1 [2]) to generate video QA pairs from EgoExo4D's keystep videos. We utilize keystep videos recorded from the ego view and all exo views, only the corresponding dense atomic action narrations are used as input to the LLM. The prompt we use aims to generate QA pairs that focus on summarizing the content of the videos, using only the dense narrations. This process results in over 50K QA pairs derived from EgoExo4D's keystep videos.