π-EGO: PROXIES AS MEDIATORS FOR POLY-MODAL-INDUCED EGOCENTRIC VIDEO REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric RGB video provides an inherently limited view of human actions, often obscuring body motion and surrounding scene context. Privileged information from complementary modalities, exocentric views, and foundation models can provide richer supervision during training, but is often unavailable or impractical at inference. To this end, we introduce **π-EGO**, a *poly-modal-induced* egocentric RGB encoder learned through **PROXYDIST**, a multi-teacher distillation framework that uses proxies as mediators to consolidate knowledge from nine heterogeneous teachers spanning viewpoints, modalities, and foundation model representations. PROXYDIST learns representation-specific *proxies* that translate heterogeneous knowledge into a homogeneous egocentric space rather than directly distilling from incompatible teacher feature spaces. Then, the proxies are merged through a learned convex initialization, followed by *Selective Proxy Distillation (SPD)*, which selectively distills from the most reliable proxies for each training sample. The resultant π-EGO achieves state-of-the-art performance across action recognition, video retrieval, and action segmentation on three challenging ego-exo benchmarks, outperforming direct multi-teacher distillation while requiring only egocentric RGB video and no additional computational overhead at inference.

1 INTRODUCTION

Understanding human actions from egocentric video is a fundamental challenge in visual perception, with applications spanning augmented reality, assistive robotics, and procedural activity analysis (Thatipelli et al., 2025; Damen et al., 2018; Sigurdsson et al., 2018). However, egocentric RGB video provides an inherently incomplete view of human activity: wearable cameras have a narrow field of view and are subject to frequent self-occlusions, often obscuring body motion and relevant scene context (Grauman et al., 2022). On the other side, complementary sensing modalities, such as depth and skeleton, can expose the geometric structure of human motion (Yan et al., 2018; Pavllo et al., 2019; Cao et al., 2019); exocentric views provide broader observations of the actor and environment; and heterogeneous foundation models (FMs) encode diverse semantic, geometric, and visual knowledge (Zhai et al., 2023; Caron et al., 2021; Yang et al., 2024; Ke et al., 2018). While recent large-scale data collection efforts demonstrate that such complementary information can be acquired as privileged supervision during training (Grauman et al., 2024; Li et al., 2024; Sener et al., 2022), requiring it at inference remains impractical for latency-sensitive egocentric applications. The challenge is therefore not merely to exploit additional information, but to internalize it into an egocentric RGB representation that no longer needs it at test time.

We call this target a poly-modal-induced egocentric representation: a single representation computed from egocentric RGB whose learning is induced by complementary knowledge across modalities, viewpoints, and heterogeneous model representations. Learning such a poly-modal-induced egocentric representation naturally suggests *multi-teacher knowledge distillation*, a paradigm extensively explored in concurrent work (Ranzinger et al., 2024; Shang et al., 2024; Wimmer et al., 2026; Belagali et al., 2025). However, existing methods largely distill homogeneous representations into a single student, sidestepping the structural challenges posed by heterogeneous supervision in the egocentric setting. First, the teachers are fundamentally heterogeneous: skeleton models operate over graph-structured architectures incompatible with video encoders (Das et al., 2020; 2021), while

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

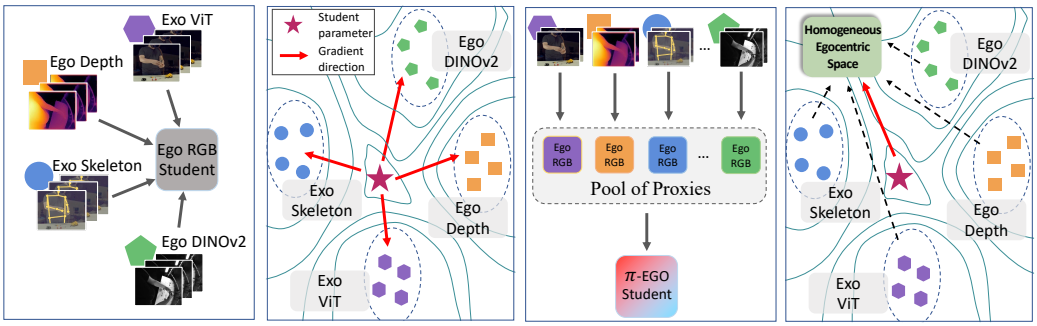

Figure 1: **(a)** Naive multi-teacher distillation with heterogeneous teachers for learning poly-modal-induced egocentric representations results in representational gaps and conflicting gradients, as illustrated in **(b)**. **(c)** In contrast, our proposed PROXYDIST framework mitigates these limitations through proxy-mediated learning to obtain π-EGO, as shown in **(d)**. Black dashed arrows illustrate the effects of this framework, shifting teachers into homogeneous proxy space.

exocentric RGB teachers observe the same activity from a different viewpoint, creating a substantial **representational gap** (Luo et al., 2025; Li et al., 2021). Second, this heterogeneity creates a difficult joint optimization problem. A single egocentric student trained to align with all teacher representations simultaneously receives distillation objectives that can induce **conflicting gradient directions** (Liu et al., 2021; Yu et al., 2020; Shi et al., 2023). Our measurements support both effects: heterogeneous teachers exhibit substantially lower pairwise representation similarity than their proxy counterparts, while direct multi-teacher distillation exhibits persistent gradient interference with the primary classification objective. Thus, simply aggregating more teachers does not resolve the central problem: the student must simultaneously translate incompatible representations and consolidate their knowledge.

Our key idea is to **translate first, consolidate second**. Instead of asking one student to reconcile heterogeneous teachers directly, we first transfer each teacher into a student-compatible egocentric **proxy**, and only then consolidate knowledge across the resulting homogeneous proxy space. Each proxy takes egocentric RGB as input, shares the architecture of the final student, and is independently distilled from one teacher. The proxy therefore acts as a mediator: it preserves knowledge induced by its corresponding modality, viewpoint, or representation while expressing that knowledge through a common egocentric architecture and input space. This decomposition separates heterogeneous knowledge translation from multi-source knowledge consolidation. Therefore, the final student learns from a collection of structurally compatible egocentric proxy models.

We instantiate this principle with PROXYDIST, a two-level proxy-mediated multi-teacher distillation framework for learning a poly-modal-induced egocentric RGB model, π-EGO. At the first level, Proxy Learning independently distills each heterogeneous teacher into its corresponding egocentric RGB proxy, transforming diverse supervisory signals into a homogeneous proxy space. At the second level, these proxies are consolidated rather than treated as uniformly trustworthy. Proxy Merging learns a convex combination of proxy parameters to initialize π-EGO from their collective knowledge, providing a favorable starting point for subsequent optimization. Selective Proxy Distillation (SPD) then performs instance-dependent reliability filtering: for each training example, it distills only from proxies that predict the correct class and are among the most confident, while suppressing proxy supervision when no proxy satisfies this criterion. Proxy Learning therefore addresses where heterogeneous knowledge is represented, while Proxy Merging and SPD address how that translated knowledge is consolidated into a single encoder. At inference, all teachers and proxies are discarded; π-EGO operates solely on egocentric RGB.

Via PROXYDIST, we train π-EGO with nine teachers spanning ego-exo viewpoints, RGB, depth, and skeleton modalities, and four FMs. π-EGO outperforms naive multi-teacher distillation across three egocentric video understanding tasks - action recognition, video retrieval, and action segmentation on three challenging ego-exo benchmarks (Sener et al., 2022; Li et al., 2024; Grauman et al., 2024).

Moreover, π-EGO generalizes across video backbone architectures, including compact models with as few as 22M parameters. We summarize our contributions as follows:

- We introduce **π-EGO**, a poly-modal-induced egocentric encoder trained via PROXYDIST, a novel proxies-as-mediators distillation framework with *nine* teachers across *ego-exo* viewpoints, *three* modalities, and *four* FMs, using representation-specific proxies as structured mediators.
- We propose three tightly coupled components in PROXYDIST: **(i)** *Proxy Learning*, which maps heterogeneous teacher supervision into homogeneous egocentric proxies, bridging modality and viewpoint gaps; **(ii)** *Selective Proxy Distillation (SPD)*, which selectively distills from the most reliable proxies per sample to mitigate gradient conflicts; and **(iii)** *Proxy Merging*, which provides a learned convex initialization for π-EGO, stabilizing SPD optimization.
- π-EGO achieves **state-of-the-art performance** across three egocentric video understanding tasks on three challenging benchmarks, demonstrating that proxy-mediated distillation yields richer, more discriminative egocentric representations than direct multi-teacher supervision.

2 RELATED WORK

2.1 EGOCENTRIC REPRESENTATION LEARNING

Over the past few years, egocentric representation learning has become a central problem in video understanding. Early works, such as EgoVLP Lin et al. (2022) and LaViLa Zhao et al. (2023), focused on learning egocentric representations from egocentric videos alone (Damen et al., 2018; Grauman et al., 2022; Sigurdsson et al., 2018; Pramanick et al., 2023). This is challenging: egocentric cameras move with the person, hands and objects frequently occlude one another, and the same action can look very different depending on the wearer and environment. More recent works augment egocentric representation learning with signals beyond the raw ego video, typically leveraging synchronized exocentric viewpoints Li et al. (2021); Xue & Grauman (2023); Quattrocchi et al. (2024); Dou et al. (2024); Luo et al. (2025) or additional modalities (Gong et al., 2023; Xu et al., 2025; Tan et al., 2023; Radevski et al., 2023). For example, ViewpointRosetta Luo et al. (2025) uses diffusion models to learn a mapping between egocentric and exocentric representation spaces, EgoDTM Xu et al. (2025) learns 3D-aware egocentric representations through distillation from a depth-modality trained teacher. These methods typically exploit a single specific auxiliary signal (viewpoint/modality). In contrast, our work aims to consolidate many heterogeneous teachers, spanning auxiliary viewpoints and modalities, into a single egocentric RGB encoder.

2.2 MULTI-TEACHER KNOWLEDGE DISTILLATION

Knowledge distillation was originally introduced as a way to compress the knowledge of a large model, or an ensemble of models, into a single deployable student (Hinton et al., 2015). This idea has since been extended to multi-teacher distillation, where a student learns from several teachers rather than a single teacher. For example, approaches like AMTML-KD Liu et al. (2020) and CA-MKD Zhang et al. (2022) adaptively distill from an ensemble of teachers operating on a single input modality, learning instance-level importance weights for each teacher. More recent work has considered a more heterogeneous form of multi-teacher distillation, where teachers are differentiated by their architecture or input modalities. For example, AM-RADIO Ranzinger et al. (2024) and Theia Shang et al. (2024) both distill multiple vision foundation models with different architectures and pretraining objectives into a single universal model. The key differences between these approaches and π-EGO are that: (1) we target poly-modal-induced egocentric representation learning that has not been explored by existing works, (2) we leverage a more diverse set of teachers spanning various architectures, modalities, and viewpoints; and (3) we selectively choose which teachers to distill from rather than treating all teachers as equally useful.

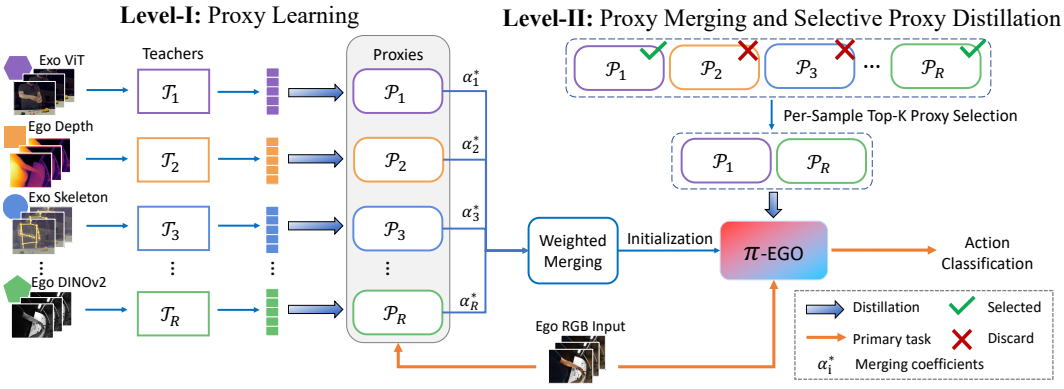

Figure 2: **Overview of PROXYDIST.** π-EGO learns a poly-modal-induced egocentric RGB encoder through a two-level proxy-mediated distillation framework. In Level-I (left), heterogeneous teachers spanning viewpoints, modalities, and foundation model representations independently supervise egocentric proxy models, converting diverse teacher signals into a homogeneous proxy space. In Level-II (right), the proxy parameters are first merged to initialize π-EGO, after which Selective Proxy Distillation (SPD) performs sample-wise reliability filtering and distills only from proxies that are correct and confident. At inference, only π-EGO operates using egocentric RGB input.

3 PROXYDIST: PROXIES-AS-MEDIATORS MULTI-TEACHER DISTILLATION FRAMEWORK

In this section, we present PROXYDIST, a multi-teacher distillation framework that uses proxies as mediators for egocentric video understanding. Let $\mathcal{D} = \{(x_i^e, \{x_i^r\}_{r=1}^R, y_i)\}_{i=1}^N$ denote the training set, where x_i^e is the egocentric video clip, x_i^r denotes the input to the r-th teacher model $\mathcal{T}_r$, and $y_i \in \mathcal{Y}$ is the ground-truth action label. In practice, x_i^r may coincide with x_i^e, correspond to its exocentric counterpart, or constitute an entirely different modality (*e.g.*, skeleton or depth), each paired with its associated teacher $\mathcal{T}_r$. Each teacher thus encodes a distinct representation, modality, or viewpoint. The objective is to learn a poly-modal-induced egocentric RGB encoder $f(\cdot)$ such that the resulting representation $f(x^e)$ subsumes the complementary knowledge of all teachers $\{\mathcal{T}_r(x^r)\}_{r=1}^R$ during training, while requiring *only* the egocentric stream x^e at inference.

Overview. PROXYDIST employs a proxies-as-mediators multi-teacher distillation strategy comprising two levels, as illustrated in Figure 2. In the *first level*, we train a set of representation-specific **Proxy** models by independently distilling knowledge from each teacher $\mathcal{T}_r$. Each proxy specializes in transferring representation- or modality- or viewpoint-specific information from its corresponding teacher, while bridging the representational gap between heterogeneous teacher representations and the egocentric input space, converting diverse supervisory signals into a homogeneous egocentric proxy space. In the *second level*, we perform Selective Proxy Distillation (SPD), which aggregates knowledge from all proxy models into the final π-EGO model, initialized via a principled proxy merging strategy. The resulting **π-EGO** effectively aggregates complementary knowledge from multiple teachers, with proxies acting as intermediaries for structured knowledge transfer. In the following subsections, we detail each level of distillation in our proposed PROXYDIST.

3.1 LEVEL-I: PROXY LEARNING

In the first distillation level, we train R proxy models $\{P_r\}_{r=1}^R$ by independently distilling feature-level knowledge from each teacher $\mathcal{T}_r$ into an egocentric student. Thus, for an egocentric video x_i^e and its time-synchronized auxiliary inputs $\{x_i^r\}_{r=1}^R$ comprising RGB, depth, and skeleton data from both egocentric and exocentric viewpoints, we extract representations via R teacher networks $\{\mathcal{T}_r\}_{r=1}^R$, each providing viewpoint- and modality-specific supervision signals.

All proxies share the same architecture but are trained with independent parameters. Since some teachers are foundation models that yield only feature embeddings rather than action logits, we adopt feature-level distillation throughout the entirety of this stage. Specifically, let $h_i, z_i = f(x_i^e)$ denote the feature embedding and action logits of the proxy student, and $h_i^r = \mathcal{T}_r(x_i^r)$ the teacher's feature

embedding. Each proxy P_r is optimized via:

$$\mathcal{L}_{\mathrm{I}}^{r} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda_{\mathrm{I}}\, D_{\cos}(h_i, h_i^r) + \lambda_{\mathrm{cls}}\ \mathrm{CE}(z_i, y_i)\right), \tag{1}$$

where $D_{\cos}$ denotes cosine distance and CE is the cross-entropy loss.

Despite being induced by heterogeneous teachers, all proxies share the same egocentric architecture. This converts heterogeneous teacher supervision into a homogeneous egocentric proxy space $\{P_r\}_{r=1}^{R}$, partitioned as $\mathcal{P} = \mathcal{P}_{\text{ego}} \cup \mathcal{P}_{\text{exo}}$ according to viewpoint. Each proxy P_r subsequently serves as a mediator in the second distillation level, reducing the modality and viewpoint gap between the original teachers and the π-EGO model.

3.2 LEVEL-II: SELECTIVE PROXY DISTILLATION (SPD)

Given the learned proxy set $\{P_r\}_{r=1}^{R}$, the goal of this level is to consolidate their complementary knowledge into a single RGB augmented egocentric model $f(\theta_\pi; \cdot)$. We first introduce a principled initialization of θ_π via proxy merging, followed by **SPD**, which selectively transfers knowledge from the most reliable proxies for each training sample.

Proxy Merging Initialization. We initialize θ_π as an optimally weighted combination of the proxy parameters. Let θ_r denote the parameters of proxy P_r. We solve for merging coefficients $\alpha^* \in \Delta^R$ that minimize the action classification loss over the training set:

$$\theta_\pi \leftarrow \theta^*_{\text{merge}}, \quad \theta^*_{\text{merge}} = \sum_{r=1}^{R} \alpha_r^*\, \theta_r, \quad \alpha^* = \arg\min_{\alpha \in \Delta^R} \frac{1}{N}\sum_{i=1}^{N} \mathrm{CE}\left(f\left(\sum_{r=1}^{R} \alpha_r \theta_r;\, x_i^e\right), y_i\right), \tag{2}$$

where $\Delta^R = \{\alpha \in \mathbb{R}^R_{\geq 0} \mid \sum_{r=1}^{R} \alpha_r = 1\}$ is the probability simplex and $f(\theta; x_i^e)$ denotes the model with parameters θ evaluated on x_i^e. This initialization places θ_π in a favorable region of the optimization landscape, providing a stable starting point for subsequent distillation.

Proposition 1 establishes that θ^*_{merge} achieves no higher classification loss than any individual proxy under local convexity, placing π-EGO in a flatter, better-generalizing region of the loss landscape (Izmailov et al., 2018; Frankle et al., 2020). This favorable initialization reduces the optimization burden of SPD, as distillation begins from a point that already encodes the consensus of all R proxy representations rather than the bias of any single one.

Proxy Selection. Rather than distilling from all proxies uniformly, SPD selects a reliable subset $\mathcal{S}_i \subseteq \{1, \ldots, R\}$ for each sample x_i^e. We adopt a correctness-filtered small-loss criterion: a proxy P_r is considered a reliable candidate only if it correctly predicts the action class, i.e., $\hat{y}_i^r = y_i$, where $\hat{y}_i^r = \arg\max_c (z_i^r)_c$. The candidate set is thus:

$$\mathcal{C}_i = \{r \in \{1, \ldots, R\} \mid \hat{y}_i^r = y_i\}. \tag{3}$$

Among candidates, proxy reliability is quantified by the cross-entropy loss $s_i^r = \mathrm{CE}(z_i^r, y_i)$, where a lower loss indicates higher predictive confidence. When $|\mathcal{C}_i| > 0$, the Top-K proxies with the lowest s_i^r are selected to form $\mathcal{S}_i$. When $\mathcal{C}_i = \emptyset$, we set $\mathcal{S}_i = \emptyset$ and skip distillation entirely for sample i, preventing the final model from absorbing erroneous supervision.

Selective Proxy Distillation. For samples where $\mathcal{S}_i \neq \emptyset$, SPD transfers knowledge from the selected proxies to π-EGO. Since all proxies are trained for action classification, they provide both feature embeddings and action logits. We therefore combine feature-level and logit-level distillation. Let $h_i^\pi, z_i^\pi = f(\theta_\pi; x_i^e)$ and $h_i^{P_r}, z_i^{P_r} = f(\theta_r; x_i^e)$ denote the feature embeddings and action logits of π-EGO and the r-th proxy P_r, respectively. The per-sample SPD loss is:

$$\mathcal{L}^{\pi}_{\text{distill}}(i) = \frac{\beta_{\text{feat}}}{|\mathcal{S}_i|}\sum_{j \in \mathcal{S}_i} D_{\cos}(h_i^\pi, h_i^{P_j}) + \frac{\beta_{\text{logit}}}{|\mathcal{S}_i|}\sum_{j \in \mathcal{S}_i} D_{\mathrm{KL}}\left(\sigma(z_i^{P_j}) \| \sigma(z_i^\pi)\right), \tag{4}$$

where $D_{\cos}$ and D_{KL} denote cosine distance and KL divergence, $\sigma(\cdot)$ is the softmax function, and $\beta_{\text{feat}}, \beta_{\text{logit}}$ are loss weights. The total Level-II objective is:

$$\mathcal{L}_{\mathrm{II}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{I}(|\mathcal{S}_i| > 0)\, \mathcal{L}^{\pi}_{\text{distill}}(i) + \beta_{\text{cls}}\, \mathrm{CE}(z_i^\pi, y_i)\right), \tag{5}$$

where β_{cls} is the classification loss weight and $\mathbb{I}(\cdot)$ is the indicator function. During inference, only π-EGO with egocentric input x^e is required: $\hat{y} = \arg\max_c \ z_c^\pi$. We note that no teacher network, exocentric stream, or proxy is needed at test time.

4 EXPERIMENTS

We evaluate π-EGO on three egocentric datasets across three video understanding tasks: action recognition, video retrieval, and action segmentation. In all experiments, π-EGO requires only egocentric RGB video at inference.

4.1 EXPERIMENTAL SETTING

Datasets. We evaluate on three publicly available ego-exo datasets: **EgoExo-Fitness** (Li et al., 2024), **Assembly101** (Sener et al., 2022), and **EgoExo4D** (Grauman et al., 2024). For EgoExo-Fitness and EgoExo4D, we follow the official evaluation splits (Li et al., 2024; Grauman et al., 2024). For Assembly101, we pair egocentric videos from the helmet-mounted *ego04* camera with exocentric videos from the frontal *exo03* camera, yielding 46,202 training and 15,307 test samples across 24 action classes. Further dataset details are provided in the Appendix.

Table 1: **Teacher pool for Level-I proxy learning.**

Modality	Viewpoint	Teacher Model	Feature Dim.
RGB	Ego	DINOv2 (Oquab et al., 2024)	1024
	Ego	SigLIP (Zhai et al., 2023)	1152
Depth	Ego	DepthAnything (Yang et al., 2024)	1024
RGB	Exo	TimeSformer (Bertasius et al., 2021)	768
	Exo	DINOv2 (Oquab et al., 2024)	1024
	Exo	Sk-Ego (Reilly et al., 2025)	512
	Exo	SigLIP (Zhai et al., 2023)	1152
Skeleton	Exo	ST-GCN (Yan et al., 2018)	256
Depth	Exo	DepthAnything (Yang et al., 2024)	1024

Implementation Details. We use TimeSformer (Bertasius et al., 2021) as the egocentric student backbone for π-EGO $f(\cdot)$ unless stated otherwise, which takes 8 frames sampled at 224×224 and produces a 768-dimensional video representation. During training, π-EGO learns from multiple teacher representations $\mathcal{T}_r$ which are detailed in Table 1. We apply a linear projection layer to align all teacher representations into the 768-dimensional student space to handle the dimensional mismatch between the student (h_i) and various teacher features (h_i^r). All proxies and the final egocentric student are trained, following TimeSformer (Bertasius et al., 2021) default configuration, for 15 epochs using SGD with a base learning rate of 0.005, momentum of 0.9, and weight decay of 1×10^{-4}. We scale the learning rate by factors of 0.1 and 0.01 at epochs 11 and 14, respectively. All models are trained with a total batch size of 8 distributed across 4 NVIDIA RTX A5000 GPUs.

In Level-I, we train 9 proxies across 6 teacher architectures and ego-exo viewpoints (Table 1), with loss weights $\lambda_{\text{I}} = 5$ and $\lambda_{\text{cls}} = 1$. In Level-II, the proxy merging coefficients α^* are optimized on the training set using Adam (Kingma & Ba, 2015) for 2 epochs with a learning rate of 0.02 and weight decay of 0.01. For proxy selection, we set $K = 1$ for EgoExo-Fitness and Assembly101, and $K = 2$ for EgoExo4D (ablated in Table 8). The Level-II loss weights are $\beta_{\text{feat}} = 5$, $\beta_{\text{cls}} = 1$ throughout, and $\beta_{\text{logit}} = 1$ for EgoExo-Fitness and EgoExo4D and 5 for Assembly101, with KL divergence temperature $\tau = 1$. All hyperparameters are selected on the respective validation sets.

We implement two controlled distillation baselines to evaluate the effectiveness of proxy-mediated knowledge transfer. i) **Multi-Teacher Distillation** directly supervises an egocentric RGB student with nine frozen teachers. Teacher-specific linear projections map the teacher's CLS token to each student's feature space and the cosine-distance losses are averaged across teachers and combined with the task-specific cross-entropy loss. ii) **Multimodal Distillation** first trains a single multimodal teacher using all available representations from the teacher pool, then distills its knowledge into an egocentric RGB student. Both baselines use the same teacher pool, student backbone, and student training schedule as π-EGO, and require only egocentric RGB video at inference.

4.2 COMPARISON WITH STATE-OF-THE-ART

Table 2 compares π-EGO with recent methods across three egocentric action recognition benchmarks under two inference protocols: *exocentric inference*, where exocentric streams are available at test

Table 2: Comparison with state-of-the-art methods on three egocentric action recognition datasets (EgoExo-Fitness, Assembly101, EgoExo4D). Dist. indicates whether the model performs distillation and Acc. indicates Top-1 accuracy.

(a) **EgoExo-Fitness.**

Method	Dist.	Acc.
Exocentric inference		
ST-GCN (Yan et al., 2018)	✗	87.5
TimeSformer (Bertasius et al., 2021)	✗	88.9
Egocentric inference		
I3D (Carreira & Zisserman, 2017)	✗	74.7
EgoVLP (Lin et al., 2022)	✗	74.7
ViFi-CLIP (Rasheed et al., 2023)	✗	81.8
π-ViT (Reilly & Das, 2024)	✓	80.1
TimeSformer (Bertasius et al., 2021)	✗	80.3
Multi-Teacher Dist. (Shang et al., 2024)	✓	81.5
Multimodal Dist. (Radevski et al., 2023)	✓	81.0
SimMMDG (Dong et al., 2023)	✓	82.2
Theia (Shang et al., 2024)	✓	81.3
π-EGO (Ours)	✓	**84.7**

(b) **Assembly101.**

Method	Dist.	Acc.
Exocentric inference		
ST-GCN (Yan et al., 2018)	✗	46.2
TimeSformer (Bertasius et al., 2021)	✗	62.7
Egocentric inference		
TSM+TA (Lin et al., 2019; Sener et al., 2020)	✗	40.5
ViFi-CLIP (Rasheed et al., 2023)	✗	46.6
π-ViT (Reilly & Das, 2024)	✓	47.8
TimeSformer (Bertasius et al., 2021)	✗	47.6
Multi-Teacher Dist. (Shang et al., 2024)	✓	48.2
Multimodal Dist. (Radevski et al., 2023)	✓	48.1
SimMMDG (Dong et al., 2023)	✓	48.7
Theia (Shang et al., 2024)	✓	48.9
π-EGO (Ours)	✓	**50.7**

(c) **Ego-Exo4D.**

Method	Dist.	Acc.
Exocentric inference		
ST-GCN (Yan et al., 2018)	✗	21.6
TimeSformer (Bertasius et al., 2021)	✗	26.0
Egocentric inference		
VI Encoder (Oord et al., 2018)	✓	40.3
EgoVLPv2 (Pramanick et al., 2023)	✓	39.1
Ego-Exo MAE (Li et al., 2021)	✓	37.2
Viewpoint Distillation (Hinton et al., 2015)	✓	38.2
TimeSformer (Bertasius et al., 2021)	✗	39.9
Multi-Teacher Dist. (Shang et al., 2024)	✓	40.6
Multimodal Dist. (Radevski et al., 2023)	✓	40.5
SimMMDG (Dong et al., 2023)	✓	40.6
Theia (Shang et al., 2024)	✓	40.7
π-EGO (Ours)	✓	**41.1**

Table 3: Performance on Video Retrieval

Method	EgoExo-Fitness		Assembly101		EgoExo4D	
	mAP	R@1	mAP	R@1	mAP	R@1
TimeSformer	47.4	71.2	22.6	41.0	16.7	32.6
Multi-Teacher Dist.	48.6	72.0	22.8	41.3	17.8	33.1
π-EGO (Ours)	**54.3**	**74.8**	**25.3**	**42.4**	**18.2**	**34.0**

Table 4: Performance on Egocentric Temporal Action Segmentation on Assembly101.

Feature Backbone (Method)	F1@10	F1@25	F1@50	Edit	Acc
TimeSformer (Ego only)	16.2	14.1	10.4	18.7	34.4
TimeSformer (Multi-Teacher Dist.)	15.3	13.2	9.8	18.4	34.2
π-EGO (Ours)	**19.6**	**16.9**	**12.3**	**19.4**	**34.7**

time, and *egocentric inference*, where only egocentric video is provided. The former serves as a privileged-view upper bound (Li et al., 2024; Sener et al., 2022), as exocentric cameras in these datasets capture global body pose and scene layout that are largely occluded in the egocentric field of view. All baselines except ST-GCN Yan et al. (2018) utilize RGB input at inference.

Under egocentric inference, π-EGO outperforms the TimeSformer backbone by +4.4%, +3.1%, and +1.2% on EgoExo-Fitness, Assembly101, and EgoExo4D respectively, demonstrating the benefit of consolidating diverse multi-teacher supervision. Compared with the strongest competing method on each dataset, SimMMDG Dong et al. (2023) on EgoExo-Fitness, Theia Shang et al. (2024) on Assembly101 and EgoExo4D, π-EGO achieves gains of +2.5%, +1.8%, and +0.4%, respectively. π-EGO also consistently surpasses naive multi-teacher distillation Shang et al. (2024) and multimodal distillation baselines Radevski et al. (2023), confirming that proxy-mediated distillation is essential for transferring heterogeneous teacher knowledge to an egocentric encoder.

The largest gains are observed on EgoExo-Fitness, where exocentric proxies $\mathcal{P}_{\text{exo}}$ are particularly strong as many actions in this dataset involve full-body motion that is inherently occluded from the egocentric viewpoint, making exocentric supervision especially informative. Conversely, on EgoExo4D, exocentric proxies are weaker, as evidenced by the baseline TimeSformer achieving only 26.0% under exocentric inference, below its egocentric counterpart. Nevertheless, π-EGO remains robust to this proxy inconsistency across viewpoints, as the adaptive selection mechanism in SPD suppresses unreliable proxies and routes supervision from the most discriminative sources available. Across all three benchmarks, π-EGO achieves state-of-the-art action recognition performance under egocentric inference.

Generalization Across Backbone Architectures. We verify that PROXYDIST is architecture-agnostic by replacing the proxy and RGB egocentric model backbone with UniFormer-S Li et al. (2022) and ViFi-CLIP Rasheed et al. (2023) (Table 5). Across both alternatives, our PROXYDIST framework consistently outperforms naive multi-teacher distillation, confirming that the gains of π-EGO are not tied to a specific video encoder. Notably, UniFormer-S, a compact 22M model yields significant classification improvements, demonstrating that our framework is equally effective for learning efficient poly-modal-induced egocentric representations, which is crucial for deployment in resource-constrained egocentric applications.

Table 5: **Backbone robustness.** π-EGO improves across diverse egocentric backbones.

Backbone	Method	EEF
TimeSformer (Bertasius et al., 2021)	baseline	80.3
	Multi-Teacher Dist.	81.5
	π-EGO (Ours)	**84.7**
UniFormer-S (Li et al., 2022)	baseline	68.4
	Multi-Teacher Dist.	69.0
	π-EGO (Ours)	**73.5**
ViFi-CLIP (Rasheed et al., 2023)	baseline	81.8
	Multi-Teacher Dist.	81.7
	π-EGO (Ours)	**83.8**

Table 6: Ablation of PROXYDIST.

Proxy Learning	Proxy Merging	SPD	EEF	A101
✗	✗	✗	80.3	47.6
✓	✗	✗	82.1	48.7
✓	✗	✓	82.3	48.9
✓	✓	✗	81.4	48.3
✓	✓	✓	**84.7**	**50.7**

Table 7: Alternative proxy merging strategies.

Merging Strategy	EEF	A101
Best Proxy	83.7	50.6
Average	83.4	50.2
Layer-level	84.2	50.6
Parameter-level	83.6	50.4
Proxy Merging (Ours)	**84.7**	**50.7**

Table 8: Strategies for Selective Proxy Distillation.

Distillation Strategy	EEF	A101
Baseline(no proxies)	80.3	47.6
All Proxies	82.8	49.9
Top-1(Ego) + Top-1(Exo)	83.2	49.5
Top-3	83.2	50.0
Top-2	84.3	50.1
Top-1	**84.7**	**50.7**

Video Retrieval & Temporal Action Segmentation. Table 3 evaluates π-EGO on video retrieval, performed by extracting features from action recognition trained backbones and computing their pairwise similarity across the test set. On EgoExo-Fitness, naive multi-teacher distillation yields only a marginal improvement of $+1.2$ mAP over the baseline, whereas π-EGO achieves a substantially larger gain of $+6.9$ mAP. This trend remains consistent across datasets, confirming that proxy-mediated distillation yields highly discriminative egocentric representations.

For temporal action segmentation on Assembly101, we extract features from three egocentric TimeSformer backbones: i) trained from scratch, ii) naive multi-teacher distillation, and iii) π-EGO, and feed them into a fixed temporal encoder Sinha et al. (2026). We evaluate the performance via F1 score, Edit distance, and frame accuracy. As shown in Table 4, π-EGO features yield the strongest performance across all metrics. Notably, naive multi-teacher distillation degrades performance on all metrics relative to the scratch baseline, confirming that directly distilling heterogeneous teachers disrupts fine-grained temporal representations. In contrast, PROXYDIST preserves the local temporal structure essential for frame-wise discrimination and boundary-sensitive segmentation.

4.3 ABLATION STUDIES AND MODEL DIAGNOSIS

We perform all the ablations and diagnosis of PROXYDIST on the EgoExo-Fitness (EEF) and Assembly101 (A101) datasets.

Effect of Each Component. Table 6 ablates the three key components of PROXYDIST: Proxy Learning, Proxy Merging Initialization, and SPD. Starting from the egocentric-only baseline, adding Level-I proxy learning followed by simultaneous distillation from all proxies yields gains of +1.8% and +1.1% on EEF and A101, respectively, highlighting the role of proxies as mediators. Then, replacing the simultaneous distillation in Level-II with SPD further improves performance by +0.2% on both datasets, demonstrating that instance-adaptive proxy selection yields more reliable supervision than distilling from all proxies indiscriminately. We also observe that proxy merging initialization provides a strong starting point for SPD, as the merged model already achieves higher accuracy than the baseline. Finally, prepending proxy merging initialization before SPD achieves 84.7% and 50.7% on EEF and A101 resulting in overall improvements of +4.4% and +3.1% over the baseline. These results confirm the contribution of each component in our PROXYDIST framework.

Alternative proxy merging strategies. Table 7 compares different strategies for initializing π-EGO before SPD. *Best Proxy* uses the strongest individual proxy as the initialization, while *Average* uniformly averages the weights of all trained proxies. We also compare against fine-grained learnable merging strategies: *Layer-level* learns separate merging weights for each layer, and *Parameter-level* learns separate merging weights for individual parameters. Although fine-grained merging offers greater flexibility, it does not improve performance in practice. Both layer-level and parameter-level merging outperform uniform averaging yet fall short of our global merging strategy, suggesting that a globally consistent combination of proxy parameters is more effective than local merging at the level of individual layers or parameter groups, likely due to the importance of maintaining parameter consistency across the full model.

Alternative SPD strategies. In Table 8, we investigate various selective distillation strategies. SPD outperforms distillation from all proxies simultaneously, suggesting that selective supervision mitigates conflicting gradients across homogeneous proxies. Enforcing viewpoint diversity via a *Top-1(Ego) + Top-1(Exo)* selection strategy degrades performance, particularly on A101, indicating that global proxy selection is preferable, i.e., not all training samples contain discriminative information from every viewpoint, and forcing viewpoint-balanced selection introduces noisy supervision into π-EGO. Among $K \in \{1, 2, 3\}$, all top-K variants outperform naive proxy distillation, demonstrating the robustness of the selection mechanism, with $K = 1$ yielding the best overall performance.

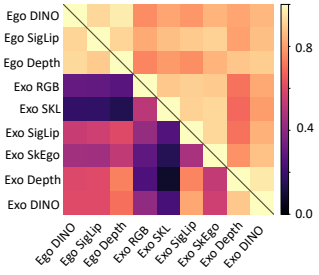

Figure 3: Teacher vs Proxy Centered Kernel Alignment scores.

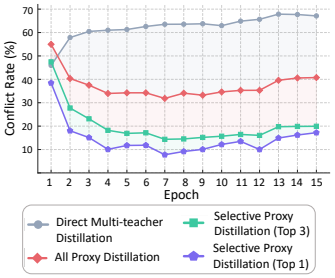

Figure 4: Gradient Conflict Rate across distillation strategies.

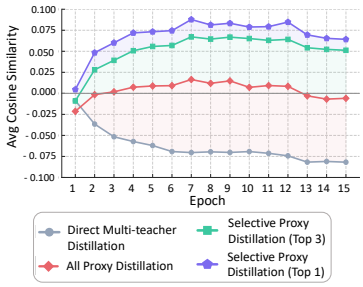

Figure 5: Cosine Similarity among teacher and proxy gradients.

5 Does ProxyDist Mitigate Representational Gap and Conflicting Gradients?

The answer is "*Yes*". We provide a model analysis to confirm ProxyDist addresses the two core failure modes of naive multi-teacher distillation: representational gap and conflicting gradients.

Representational Gap. To quantify the representational gap among the R teachers, we visualize their pairwise linear Centered Kernel Alignment (CKA) Kornblith et al. (2019) in Figure 3. The lower triangle displays teacher-pair similarities and the upper displays proxy-pair similarities on EgoExo-Fitness. The substantially higher pairwise CKA among proxies than among teachers confirms that Level-I proxy learning projects heterogeneous teacher representations into a homogeneous egocentric proxy space, directly alleviating the representational gap that impedes naive multi-teacher distillation.

Conflicting Gradients. Figures 4 and 5 report the gradient conflict rate defined as the fraction of distillation gradients that oppose the classification gradient, i.e., $\cos(\nabla_{\text{cls}}, \nabla_{\text{kd}}) < 0$, and the average cosine similarity across training epochs, respectively. Direct multi-teacher distillation exhibits severe gradient interference, confirming that heterogeneous teacher supervision actively opposes the primary task objective. Distilling from all proxies simultaneously reduces the conflict rate to 42–50% and shifts the cosine similarity toward zero, indicating partial mitigation. SPD further suppresses conflicts by routing supervision exclusively through reliable proxies, yielding consistently cooperative distillation and classification gradients. Together, these results confirm that proxy learning and SPD collectively transform a conflicting multi-teacher optimization landscape into a coherent one.

6 Conclusion

We presented **π-Ego**, a poly-modal-induced egocentric RGB encoder that consolidates perceptual knowledge across viewpoints, modalities, and foundation model representations into a single model. At the core of π-Ego lies ProxyDist, a proxies-as-mediators multi-teacher distillation framework: representation-specific proxies first bridge the representational gap between heterogeneous teachers, and reliability-guided selective distillation then integrates knowledge into a coherent poly-modal-induced egocentric representation. The resulting egocentric encoder sees further, knows more, and generalizes better without requiring more than a single egocentric camera at inference.

While we hope π-Ego inspires broader exploration into systematically harvesting and reconciling diverse supervisory signals for richer egocentric perception, the current framework has an important limitation. Specifically, π-Ego has a relatively heavy training cost, which arises primarily from learning the R proxies. Two directions could reduce this overhead. First, since the proxies serve only as mediators, they only need to provide sufficiently informative supervision rather than fully converge, so their training epochs can be reduced. Second, the current small-loss heuristic for proxy selection could be replaced by a learned mechanism that weighs proxy reliability conditioned on the input and training state; such reliability estimates could further guide pruning of the proxy pool. Designing such an adaptive selection and pruning strategy is non-trivial, and we leave it as a promising direction for future work.

7 AI USE STATEMENT

In this work, we used generative AI tools to edit the research paper to improve readability. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results, design or provide feedback on research methodology or experiments, or assist with translation. Additionally, we used generative AI tools for developing a user-friendly running script. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

8 REPRODUCIBILITY STATEMENT

We provide details on implementation of π-EGO in Sec 4.1, including the network architecture, teacher models, proxy training procedure, hyperparameters, and training schedules. Details of the datasets are provided in Appendix B. We will publicly release all code, pretrained model weights, and training and evaluation scripts upon acceptance.

REFERENCES

Varun Belagali, Saarthak Kapse, Pierre Marza, Srijan Das, Zilinghan Li, Sofiène Boutaj, Pushpak Pati, Srikar Yellapragada, Tarak Nath Nandi, Ravi K Madduri, Joel Saltz, Prateek Prasanna, Stergios Christodoulidis, Maria Vakalopoulou, and Dimitris Samaras. Ticon: A slide-level tile contextualizer for histopathology representation learning, 2025. URL https://arxiv.org/abs/2512.21331.

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)*, July 2021.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Conference on Computer Vision and Pattern Recognition*, pp. 4724–4733. IEEE, 2017.

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.

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*, pp. 72–90. Springer, 2020.

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*, pp. 1–1, 2021.

Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. SimMMDG: A simple and effective framework for multi-modal domain generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zi-Yi Dou, Xitong Yang, Tushar Nagarajan, Huiyu Wang, Jing Huang, Nanyun Peng, Kris Kitani, and Fu-Jen Chu. Unlocking exocentric video-language data for egocentric video representation learning, 2024. URL https://arxiv.org/abs/2408.03567.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning (ICML)*, 2020.

Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. Mmg-ego4d: Multi-modal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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 Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, 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 Puentes, 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 Arbelaez, 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*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zachary Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, María Escobar, Cristhian Forigua, Abrham Kahsay Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Dutt Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J. Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh K. Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Ángela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anushk Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Romy Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, W. Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Ze Yu, Shengxin Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, J. F. Zhuo, Pablo Arbeláez, Gedas Bertasius, David J. Crandall, Dima Damen, Jakob Julian Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard A. Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Uncertainty in Artificial Intelligence (UAI)*, 2018.

Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Learning clip representations for skeleton-based 3d action recognition. *IEEE Transactions on Image Processing*, 27(6):2842–2855, June 2018. ISSN 1941-0042. doi: 10.1109/TIP.2018.2812099.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Jiao Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:12581–12600, 2022. URL `https://api.semanticscholar.org/CorpusID:246240170`.

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)*, pp. 10995–11005, 2021.

Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Egoexo-fitness: towards egocentric and exocentric full-body action understanding. In *European Conference on Computer Vision*, pp. 363–382. Springer, 2024.

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.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18873–18885. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/9d27fdf2477ffbff837d73ef7ae23db9-Paper.pdf.

Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 2020.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. 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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5285–5297, 2023.

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.

Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5213–5224, 2023.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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)*, June 2024.

Dominick Reilly, Manish Kumar Govind, Le Xue, and Srijan Das. From my view to yours: Ego-augmented learning in large vision language models for understanding exocentric daily living activities, 2025.

Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *European conference on computer vision*, pp. 154–171. Springer, 2020.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jinghuan Shang, Karl Schmeckpeper, Brandon B May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. *arXiv preprint arXiv:2407.20179*, 2024.

Guangyuan Shi, Qimai Li, Wenlong Zhang, Jiaxin Chen, and Xiao-Ming Wu. Recon: Reducing conflicting gradients from the root for multi-task learning. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=ivwZO-HnzG_.

Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

Arkaprava Sinha, Monish Soundar Raj, Pu Wang, Ahmed Helmy, Hieu Le, and Srijan Das. Ms-temba: Multi-scale temporal mamba for understanding long untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9815–9826, June 2026.

Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. *arXiv preprint arXiv:2301.02217*, 2023.

Anirudh Thatipelli, Shao-Yuan Lo, and Amit K. Roy-Chowdhury. Egocentric and exocentric methods: A short survey. *Computer Vision and Image Understanding*, 257:104371, 2025. ISSN 1077-3142.

Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. Anyup: Universal feature upsampling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Boshen Xu, Yuting Mei, Xinbi Liu, Sipeng Zheng, and Qin Jin. Egodtm: Towards 3d-aware egocentric video-language pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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.

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.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5434–5445. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/3fe78a8acf5fda99de95303940a2420c-Paper.pdf.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2023.

Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Yue Zhao, Ishan Misra, Philipp Krahenbuhl, and Rohit Girdhar. Learning video representations from large language models. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6586–6597, 2023.

APPENDIX

The appendix is categorized into the following parts:

- Section A: Proxy Merging Proposition
- Section B: Detailed Data Description
- Section C: Proxy Performances
- Section D: Analysis of Proxy Selection
- Section E: Analysis of Viewpoint Selection
- Section F: Analysis of Proxy Merging
- Section G: Action Analysis

A PROXY MERGING PROPOSITION

Proposition 1 (Proxy Merged Initialization as a Loss Upper Bound for π-EGO)**.** *Let* $\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CE}(f(\theta; x_i^e), y_i)$ *denote the classification loss. Assume* $\mathcal{L}$ *is convex in* θ *in a neighborhood containing* $\{\theta_r\}_{r=1}^{R}$*. Then for any* $\alpha \in \Delta^R$*:*

$$\mathcal{L}\left(\sum_{r=1}^{R} \alpha_r \theta_r\right) \leq \sum_{r=1}^{R} \alpha_r \, \mathcal{L}(\theta_r), \tag{6}$$

and consequently the optimally merged initialization satisfies: $\mathcal{L}(\theta^*_{merge}) \leq \min_r \, \mathcal{L}(\theta_r)$.

The first inequality follows directly from Jensen's inequality Boyd & Vandenberghe (2004) applied to the convex loss $\mathcal{L}$. The second follows by noting that α^* is chosen to minimize $\mathcal{L}(\sum_r \alpha_r \theta_r)$, and the degenerate solution $\alpha_r = 1$ for any single r is feasible in Δ^R, so the optimum is no worse than the best individual proxy.

B DETAILED DATA DESCRIPTION

EgoExo-Fitness Li et al. (2024) is a full-body action understanding dataset containing synchronized egocentric and exocentric videos of fitness activities. It consists of 12 action categories and approximately 32 hours of video. Following the official split, we use 3,522 training samples and 912 test samples, with action sequences ranging from 10 to 30 seconds in duration.

Assembly101 Sener et al. (2022) is a large-scale procedural activity dataset comprising 167 hours of video, recording subjects performing object assembly tasks from multiple viewpoints. In our experiments, we pair the egocentric videos from the helmet-mounted *ego04* camera with the corresponding exocentric videos in front *exo03* camera. The resulting split contains 46,202 training samples and 15,307 test samples over 24 action classes.

EgoExo4D Grauman et al. (2024) is a large-scale egocentric-exocentric video dataset covering a diverse range of skilled human activities. For each sample, we pair the egocentric video with the exocentric view annotated as the best view by human annotators. The dataset split used in our experiments contains 30,660 training samples and 9,356 test samples across 665 action classes.

C PROXY PERFORMANCES

Foundation models (FMs) are not directly optimized for action recognition; their role is to provide complementary representations that improve the egocentric student. To directly evaluate each teacher representation, we train a linear action-classification head on top of the corresponding frozen FM. Table 9 compares this linear-probing performance with that of the corresponding learned proxy after Level-I distillation.

Table 9: Comparison between frozen foundation models evaluated using a linear classification head and their corresponding learned proxies after Level-I distillation. Results are reported on EgoExo-Fitness (EEF), Assembly101 (A101), and EgoExo4D (EE4D).

Modality/ Representation	Viewpoint	Model	Frozen FM (Linear Probe)			Learned Proxy / Model		
			EEF	A101	EE4D	EEF	A101	EE4D
Base Ego Model								
RGB	Ego	TimeSformer (Bertasius et al., 2021)		–		80.3	47.6	39.9
Ego Proxies ($\mathcal{P}_{\text{ego}}$)								
Depth	Ego	DepthAnything (Yang et al., 2024)	38.94	30.61	21.08	80.3	48.2	40.1
DINOv2	Ego	DINOv2 (Oquab et al., 2024)	40.52	30.49	23.92	80.9	48.4	40.2
SigLIP	Ego	SigLIP (Zhai et al., 2023)	32.40	29.35	21.47	80.7	48.4	40.6
Exo Proxies ($\mathcal{P}_{\text{exo}}$)								
RGB	Exo	TimeSformer (Bertasius et al., 2021)		–		81.9	49.5	40.0
Skeleton	Exo	ST-GCN (Yan et al., 2018)		–		81.9	48.4	40.3
Depth	Exo	DepthAnything (Yang et al., 2024)	39.05	31.49	10.07	80.3	48.5	39.9
DINOv2	Exo	DINOv2 (Oquab et al., 2024)	45.93	32.04	12.81	80.7	48.2	39.5
RGB	Exo	Sk-Ego (Reilly et al., 2025)		–		81.9	48.3	40.2
SigLIP	Exo	SigLIP (Zhai et al., 2023)	41.18	25.19	11.71	80.3	48.0	40.6
Final Model								
RGB	Ego	**π-EGO (Ours)**		–		**84.7**	**50.7**	**41.1**

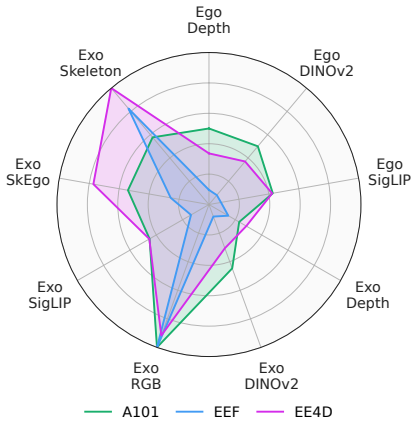

Figure 6: **Selective Proxy Distillation Statistics.**

The frozen FMs exhibit substantially lower action-recognition performance because they are not directly optimized for this task. In contrast, their corresponding learned proxies achieve considerably stronger performance after Level-I distillation, demonstrating that the complementary FM representations can be effectively transferred to the action-recognition task. Nevertheless, not all learned proxies outperform the base ego model, motivating the need for proxy selection. The final π-EGO model surpasses every individual proxy, corroborating the effectiveness of our hierarchical distillation framework.

D ANALYSIS OF PROXY SELECTION.

In Figure 6, we analyze the selection frequency of each proxy during the Level-II selective proxy distillation (SPD). First, every proxy is selected to some extent across all benchmarks, indicating that SPD does not collapse to a single teacher. Unsurprisingly, we find that the routing distributions are not consistent across benchmarks, and that proxy selection is strongly correlated with the distinct dataset characteristics. On **EgoExo-Fitness**, the SPD primarily selects *Exo Skeleton* and *Exo RGB*. This is consistent with the nature of the dataset, which consists largely of full-body fitness motions: exocentric and skeleton-based proxies provide supervision for body pose and motion patterns that are often only partially visible from a head-mounted camera. In **Assembly101**, the selection is more diverse, assigning a larger weight to egocentric proxies. This suggests that, compared to EgoExo-Fitness, more actions in Assembly101 rely on egocentric cues such as local hand-object interactions and subtle manipulation dynamics. On **EgoExo4D**, there is a notable increase in the

selection of the *Exo SkEgo* proxy. We attribute this to the high-resolution, spatially localized nature of the benchmark: action-relevant regions can occupy only a small part of the full frame, and the SkEgo crop provides a more focused view of these regions.

E Analysis of Viewpoint Selection.

Exocentric videos offer a broader spatial context Luo et al. (2025), since global actions and body motions are often invisible from the ego view, which motivated us to use them as privileged input. We report accuracy with ego-only and exo-only teachers in Table 10:

Table 10: Results under different viewpoint access.

Supervision \ Dataset	EgoExo-Fitness	Assembly101	EgoExo4D
Ego proxies only	83.09	49.12	40.67
Exo proxies only	84.33	50.11	41.03
All proxies (Ours)	84.7	50.7	41.1

Exo proxies contribute most of the gain, as expected from their richer spatial context, but the exo-only model still underperforms the full model: since inference is in the egocentric view, ego teachers remain complementary and guide the student (e.g., ego_depth guides spatial-layout reasoning, ego_dino the semantics of interacting objects). Moreover, the gain is not merely from exocentric access: π-EGO outperforms direct multi-teacher distillation with the same nine frozen teachers in Table 2, showing the improvement stems from the framework.

F Analysis of Proxy Merging

Table 11: **Effect of different proxy initialization strategies before and after SPD on EgoExo-Fitness (EEF) and Assembly101 (A101).**

Method	EEF		A101	
	Before SPD	After SPD	Before SPD	After SPD
Best Proxy	81.9	83.7	49.5	50.6
Average	79.9	83.4	42.9	50.2
Layer-level	80.9	84.2	43.5	50.6
Parameter-level	79.8	83.6	43.4	50.4
Model-level	81.4	84.7	48.3	50.7

Table 11 compares action classification performance of π-EGO under different proxy merging initialization strategies, before and after SPD, on EgoExo-Fitness and Assembly101. Across all initialization strategies, SPD consistently improves performance, confirming the importance of the second distillation level. Notably, the best-performing initialization prior to SPD, i.e., the single Best Proxy, does not yield the strongest post-SPD accuracy. Instead, the model soup, i.e., the learned convex combination of all proxy parameters, provides the most effective initialization for SPD. This suggests that the advantage of proxy merging lies not in its standalone classification performance but in its ability to place π-EGO in a well-conditioned region of the loss landscape that is maximally amenable to subsequent distillation.

G Action Analysis

As shown in Table 12 and Table 13, `exo_rgb` is the most frequently selected proxy on both datasets, with a sample-weighted selection rate of 0.349 on EEF and 0.182 on A101, followed by `exo_skl`. However, the class-level separation between the top-ranked and second-ranked proxies differs substantially across datasets. The average per-class gap between the top-1 and top-2 proxies is 0.186 on EEF, but only 0.069 on A101. This suggests that EEF classes often have a more clearly

Table 12: Per-class top-1 proxy selection rates on EEF (12 classes)

class_name	none	exo_rgb	exo_skl	exo_siglip	ego_siglip	exo_skego	ego_depth	exo_depth	exo_dino	ego_dino
Kneeling pushing-ups	0.053	0.139	0.172	0.053	0.053	0.139	0.013	**0.179**	0.106	0.093
Push-ups	0.000	**0.500**	0.400	0.000	0.025	0.042	0.008	0.017	0.000	0.008
Kneeling Torso Twist	0.046	**0.431**	0.218	0.030	0.005	0.056	0.117	0.041	0.036	0.020
Knee Raise & Abd. Contract	0.055	0.290	**0.428**	0.014	0.028	0.048	0.055	0.034	0.021	0.028
Shoulder Bridge	0.000	**0.987**	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Sit-ups	0.057	0.172	**0.261**	0.070	0.013	0.108	0.076	0.146	0.045	0.051
Leg Reverse Lunge	0.038	0.106	**0.477**	0.068	0.068	0.114	0.023	0.045	0.053	0.008
Leg Lunge With Knee Lift	0.118	**0.300**	0.291	0.073	0.036	0.109	0.036	0.018	0.009	0.009
Sumo Squat	0.112	**0.388**	0.233	0.017	0.017	0.198	0.009	0.009	0.000	0.017
Jumping Jacks	0.042	0.242	**0.347**	0.126	0.016	0.153	0.016	0.005	0.021	0.032
High Knee	0.054	0.264	**0.271**	0.085	0.078	0.078	0.008	0.054	0.054	0.054
Clap Jacks	0.011	0.345	**0.414**	0.029	0.023	0.063	0.029	0.057	0.011	0.017

Table 13: Per-class top-1 proxy selection rates on A101 (24 classes).

class_name	none	exo_rgb	exo_skl	exo_siglip	ego_siglip	exo_skego	ego_depth	exo_depth	exo_dino	ego_dino
pick up	0.049	**0.256**	0.125	0.089	0.079	0.104	0.079	0.042	0.076	0.102
put down	0.051	**0.204**	0.122	0.107	0.106	0.103	0.091	0.045	0.076	0.094
inspect	0.186	**0.136**	0.108	0.080	0.066	0.092	0.085	0.052	0.091	0.105
rotate	0.307	**0.123**	0.085	0.056	0.049	0.090	0.088	0.037	0.083	0.082
unscrew	0.042	**0.196**	0.124	0.100	0.100	0.123	0.124	0.044	0.064	0.085
position	0.107	0.139	0.081	0.075	0.086	**0.143**	0.125	0.048	0.088	0.108
screw	0.048	**0.250**	0.078	0.099	0.087	0.088	0.128	0.042	0.087	0.091
remove	0.187	0.125	**0.138**	0.064	0.067	0.098	0.090	0.044	0.085	0.102
position screw on	0.134	**0.141**	0.086	0.104	0.084	0.088	0.111	0.039	0.107	0.106
remove screw from	0.201	0.115	0.117	0.092	0.068	0.071	0.083	0.043	**0.141**	0.071
pass	0.165	**0.142**	0.089	0.059	0.067	0.123	0.086	0.040	0.130	0.098
tilt up	0.291	**0.143**	0.041	0.058	0.082	0.057	0.102	0.042	0.087	0.097
attempt to position	0.413	0.084	**0.097**	0.036	0.025	0.088	0.089	0.035	0.048	0.085
push	0.336	0.063	0.098	0.076	0.096	**0.102**	0.061	0.054	0.078	0.036
tilt down	0.224	**0.179**	0.116	0.078	0.068	0.108	0.046	0.041	0.079	0.059
pull	0.370	0.082	**0.123**	0.042	0.062	0.071	0.076	0.027	0.120	0.029
attempt to remove	0.595	0.067	0.031	0.029	0.012	0.084	0.061	0.006	0.027	**0.088**
attempt to pick up	0.754	0.040	0.022	0.025	0.020	**0.047**	0.025	0.017	0.017	0.032
clap	0.007	**0.883**	0.047	0.010	0.005	0.015	0.005	0.007	0.012	0.007
attempt to unscrew	0.515	0.095	0.075	0.037	0.050	0.025	0.079	0.008	0.017	**0.100**
attempt to put down	0.988	**0.012**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
spin	0.452	**0.170**	0.064	0.053	0.027	0.059	0.037	0.027	0.032	0.080
attempt to screw	0.951	**0.019**	0.010	0.000	0.000	0.010	0.000	0.000	0.010	0.000
shake	0.783	**0.075**	0.000	0.025	0.000	0.017	0.008	0.000	0.042	0.050

preferred proxy, whereas A101 exhibits a flatter proxy distribution, where several proxies can have similar selection rates for the same class.

Interestingly, `exo_depth` shows a highly class-specific behavior. Although it has the lowest aggregate selection rate among all proxies (0.042 on A101 and 0.052 on EEF), it is the most frequently selected proxy for *Kneeling push-ups* on EEF, with a selection rate of 0.179, exceeding the second-ranked proxy, `exo_rgb`, at 0.139. This indicates that depth can provide useful supervision for specific classes where geometric cues, such as body-to-floor distance, are discriminative. However, its low aggregate selection rate also suggests that this benefit is not broadly shared across classes.

The behavior of ego-side proxies also differs across the two datasets. On EEF, all ego-side proxies have relatively low selection rates, around 0.03, suggesting that third-person supervision is more reliable for large whole-body fitness motions. In contrast, on A101, `ego_dino` and `ego_depth` reach selection rates of approximately 0.091 and rank among the middle group of proxies. Each of them is also the most frequently selected proxy for at least one *attempt-to-** class. This difference is consistent with the task characteristics of A101: fine-grained assembly actions are often hand- and object-centric, and egocentric observations may capture local manipulation details that are less visible from exocentric views.