Viewpoint Rosetta Stone: Unlocking Unpaired Ego-Exo Videos for View-invariant Representation Learning

Mi Luo[1] Zihui Xue[1] Alex Dimakis[2, 3] Kristen Grauman[1]
[1]The University of Texas at Austin [2]UC Berkeley [3]Bespoke Labs

Abstract

Egocentric and exocentric perspectives of human action differ significantly, yet overcoming this extreme viewpoint gap is critical in augmented reality and robotics. We propose VIEWPOINTROSETTA*, an approach that unlocks large-scale unpaired ego and exo video data to learn clip-level viewpoint-invariant video representations. Our framework introduces (1) a diffusion-based Rosetta Stone Translator (RST), which, leveraging a moderate amount of synchronized multi-view videos, serves as a translator in feature space to decipher the alignment between unpaired ego and exo data, and (2) a dual encoder that aligns unpaired data representations through contrastive learning with RST-based synthetic feature augmentation and soft alignment. To evaluate the learned features in a standardized setting, we construct a new cross-view benchmark using Ego-Exo4D, covering cross-view retrieval, action recognition, and skill assessment tasks. Our framework demonstrates superior cross-view understanding compared to previous view-invariant learning and ego video representation learning approaches, and opens the door to bringing vast amounts of traditional third-person video to bear on the more nascent first-person setting.* [1]

1. Introduction

Human perception of action is profoundly view-invariant. No matter the angle from which we observe an action—whether we're viewing a person from a third-person (exocentric) perspective or experiencing it firsthand (egocentric)—our understanding of that action remains stable. This capacity allows us to recognize, interpret, and respond to complex movements and interactions across diverse perspectives. However, existing computer vision models struggle to replicate this ego-exo[2] view-invariant understanding of action. The extreme variations in visual appearance of the two views make it difficult for models to generalize across the two viewpoints (see Figure 1, left).

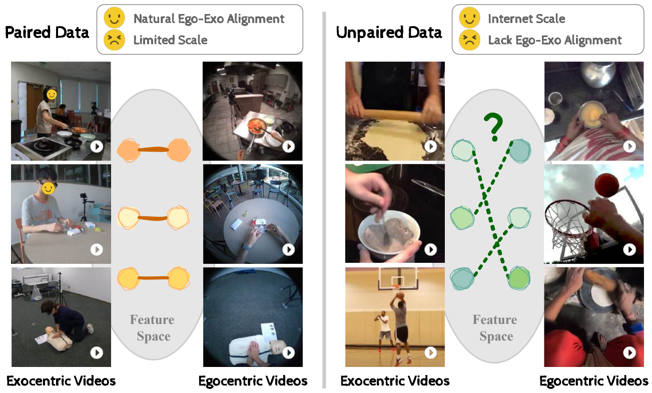

Figure 1. Paired data is ideal for ego-exo view-invariant representation learning due to its perfect synchronization, but it is costly to collect. We explore leveraging both paired and unpaired data, taking advantage of the greater scale of unpaired videos. The key question is how to discover meaningful links within unpaired data and how to effectively align ego and exo representations.

We take a step toward scaling ego-exo view-invariant representation learning, which can benefit a range of applications that require cross-view understanding. View-invariant representations are essential to enable cross-view retrieval in AR, such as automatically retrieving a relevant exo video from YouTube that aligns with the user's current cooking setup and tools as viewed from their wearable camera, showing a chef performing the same task and providing real-time guidance. Similarly, robot learning from video demands being able to interchangeably understand how an action looks when watching a human subject versus how it looks from the first-person view when performing dexterous manipulation. In addition, we see practical potential for transferring knowledge from the exocentric domain to the egocentric domain—resources for exocentric video are far more extensive, while egocentric datasets (e.g., Ego4D [16], Ego-Exo4D [17], EPIC-Kitchens [9, 10]) are relatively new and not yet at the Internet scale.

[1] Project webpage: https://vision.cs.utexas.edu/projects/ViewpointRosetta/

[2] We use "ego" to refer to egocentric (first-person) and "exo" to refer to exocentric (third-person) perspectives.

Previous approaches to learn ego-exo view invariance have relied on limited amounts of paired multi-view video data [3, 17, 36, 37, 49]. However, collecting synchronized ego-exo video is both costly and logistically challenging, requiring specialized equipment to capture footage from both perspectives simultaneously. This remains an obstacle to developing models that bridge the gap between these viewpoints effectively. Early attempts to learn view-invariant features from unpaired data lack adaptability to diverse, unstructured data due to their reliance on limited forms of alignment, like temporal cues [48] or partially overlapping semantics [42].

Our goal is to unlock large-scale *unpaired* ego and exo video data for view-invariant video representation learning. Unpaired data offers greater scalability and diversity by eliminating the need for synchronized video capture, while also providing a broader range of semantic information, such as varied settings, objects, and interactions, which should help the model learn more generalized and robust representations. See Figure 1.

Our key insight is to leverage a moderate amount of paired, time-synchronized ego-exo video as a "Rosetta Stone"[3] to 1) master the ego-exo link and then 2) interpret complex relationships within unpaired ego and exo data. Specifically, we introduce a Rosetta Stone Translator (RST) based on a diffusion model [30] that learns the mapping between ego and exo viewpoints in feature space. Given an unpaired ego sample, the translator generates its exo counterpart. RST is trained on time-synchronized video data, providing a strict temporal and geometric mapping between the two viewpoints. It learns a precise mapping between the ego and exo views by capturing how each perspective represents the same scene—such as detailed hand-object interactions in the ego view and broader spatial context in the exo view. Having learned this relationship, the RST is then equipped to generate features for the alternate view that integrate both detail and context.

The synthesized features from the RST are then used to establish connections between real-world, unpaired samples. To transform these established connections into a view-invariant representation, we train a video-text dual encoder on both paired and RST-aligned unpaired data, reformulating the classic contrastive loss to incorporate the synthetic data generated by the RST as an augmentation. Additionally, we introduce a notion of "soft alignment" to handle the fact that most unpaired data are not perfectly matched at the semantic level. This soft alignment allows for minor discrepancies in the paired samples, focusing on capturing essential shared information rather than exact visual matches.

To evaluate the learned view-invariant representations, we establish an ego-exo cross-view understanding benchmark using the Ego-Exo4D [17] dataset. This benchmark covers a variety of downstream tasks, including cross-view recognition, cross-view retrieval, and cross-view skill assessment. Across these tasks, our model consistently outperforms previous methods for view-invariant learning [17, 37, 42] as well as state-of-the-art video representation learning models [24, 51]. Our work introduces a novel perspective by tapping into the vast potential of extensive-scale unpaired videos, setting new standards for ego-exo view-invariant learning in future research.

2. Related Work

Bridging Egocentric and Exocentric Views Ego and exo view videos inherently differ, presenting unique challenges and perspectives for video understanding. There have been several attempts to jointly study these two views, introducing ego-exo datasets [14, 17, 21, 35, 38] for this purpose and addressing various challenges, such as person localization [1, 2, 15, 43, 47], video summarization [19], and 3D pose estimation [41].

To advance cross-view understanding between these two disparate views, several studies [3, 17, 36, 37, 42, 48, 49] have focused on ego-exo view-invariant feature learning, which aims at pushing relevant ego-exo views closer and irrelevant farther. However, existing approaches face stringent data constraints: they are either confined to using strictly paired ego-exo datasets [3, 17, 36, 37, 49] or utilize unpaired datasets limited to hundreds of hours [42, 48], and aim to improve models for egocentric video tasks.

This is notably inadequate when compared to the extensive scale of Internet-scale video collections, like Ego4D [16] with 3,000 hours of videos and HowTo100M [28] featuring 136 million video clips. In contrast, our work innovatively unlocks these large-scale, in-the-wild video datasets, demonstrating new potential in ego-exo cross-view understanding.

Another line of work explores synthesizing unseen exo views from ego views (or vice versa) in the pixel space, for the sake of visualization [8, 17, 25–27]. Though our approach includes view translation (in the feature space), our purpose is orthogonal: to form a bridge between the viewpoints for feature learning.

Egocentric Video Representation Learning Building strong representations for egocentric videos is crucial, enhancing various downstream tasks. One area of research focuses on video-text representation learning [4, 24, 32, 51], leveraging narrations accompanying videos for training, to improve video-text tasks such as natural language grounding [39]. Another line of work [13, 23, 40] directly addresses egocentric representation learning, by utilizing auxiliary ego signals from exo videos [23], or employing MAE training on ego videos [40] to improve downstream egocentric tasks such as action recognition. Our work concen-

[3] https://en.wikipedia.org/wiki/Rosetta_Stone

trates on *ego-exo cross-view understanding*, aiming to link ego and exo-view videos across tasks such as cross-view retrieval. We advocate an orthogonal perspective—ego-exo view-invariance in the representation learning process, which we will show to be superior for cross-view understanding compared with prior video-text or vision-centric non view-invariant approaches.

Generative Models for Representation Learning It is widely believed that strong generative capabilities often signal the potential for learning robust representations. Previous research has demonstrated that diffusion models can capture meaningful representations for image recognition tasks [7, 22, 44, 50]. We build on this insight and introduce a new application of diffusion representations to facilitate video representation learning. This process enables the synthetic creation of pseudo-paired exo features from a single ego video, which is particularly valuable for learning ego-exo view-invariant features given the scarcity and limited scale of existing paired ego-exo videos.

3. Methodology

We first define the problem formally (Sec. 3.1), then introduce the Rosetta Stone Translator (Sec. 3.2). Next, building on that translator, we explain how we align the unpaired video data (Sec. 3.3) and then train using our newly pseudo-aligned data with contrastive learning (Sec. 3.4).

3.1. Problem Setup

Prior approaches to learning view-invariant video representations have relied on training contrastive models using **paired ego-exo videos**, either time-sychronized [17] or pseudo-synchronized videos [37]. Let $\mathcal{D}_{\text{pair}}$ denote a dataset of paired videos, where each pair consists of an ego view and a synchronized exo view of the same scene: $\mathcal{D}_{\text{pair}} = \{(v_{\text{ego}}^{(i)}, v_{\text{exo}}^{(i)})\}_{i=1}^{N_{\text{pair}}}$, where $v_{\text{ego}}^{(i)}$ and $v_{\text{exo}}^{(i)}$ are the ego and exo views, respectively, of the i-th synchronized pair.

The training objective commonly employs a contrastive loss, such as InfoNCE [29], to bring representations of synchronized pairs closer together in feature space, while pushing non-synchronized pairs further apart. For each mini-batch of N paired samples drawn from $\mathcal{D}_{\text{pair}}$, a video encoder $f_\theta : V \to \mathbb{R}^d$ is trained to maximize the similarity between the ego and exo features of the same sample, while minimizing similarity with other samples within the batch. The contrastive loss $\mathcal{L}_{\text{c}}$ for a batch of paired samples is defined as:

$$\mathcal{L}_{\text{c}}(u, v) = -\log \frac{\exp(\text{sim}(u, v)/\tau)}{\sum_{x \in \mathcal{B}} \exp(\text{sim}(u, x)/\tau)}$$

where $u = f_\theta(v_{\text{ego}}^{(i)})$, $v = f_\theta(v_{\text{exo}}^{(i)})$, $\text{sim}(\cdot, \cdot)$ denotes a similarity metric, such as cosine similarity. $\mathcal{B}$ represents all samples in the batch. τ is a temperature parameter that controls the concentration of the similarity distribution.

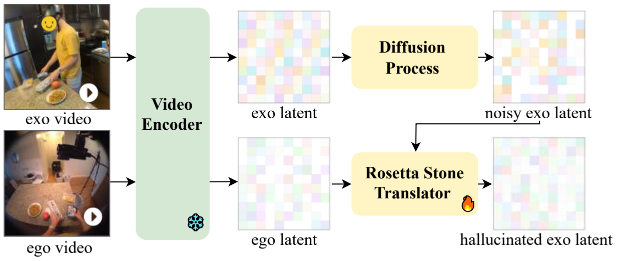

Figure 2. Training of the Rosetta Stone Translator (RST). Leveraging synchronized ego and exo videos, we extract features with a frozen video encoder f_ϕ. The RST is trained to predict exo features from ego ones, using a denoising network to reverse the diffusion process.

To enable view-invariant learning at a large scale, we expand the setup to include both synchronized paired videos and unpaired in-the-wild videos. The unpaired video data consists of two separate datasets: an ego-only dataset, $\mathcal{D}_{\text{ego}} = \{v_{\text{ego}}^{(i)}\}_{i=1}^{N_{\text{ego}}}$, and an exo-only dataset, $\mathcal{D}_{\text{exo}} = \{v_{\text{exo}}^{(j)}\}_{j=1}^{N_{\text{exo}}}$, where in practice $N_{\text{exo}} > N_{\text{ego}}$ and both N_{ego} and N_{exo} are large. These datasets capture various scenes and actions from different viewpoints, with videos recorded independently, making them unpaired in both temporal and contextual alignment.

Our goal is to train a **view-invariant video encoder** f_θ using both paired data $\mathcal{D}_{\text{pair}}$ and unpaired data $\mathcal{D}_{\text{ego}}$ and $\mathcal{D}_{\text{exo}}$, such that synchronized ego-exo pairs are embedded close together in the feature space, reinforcing viewpoint invariance for paired samples. Unpaired videos with similar visual and semantic content across ego and exo datasets are also encouraged to be close in the feature space. We assume that both paired and unpaired data are accompanied by text narration, which can be easily obtained through ASR or state-of-the-art video captioning models [5, 45]. In our context, language plays a supplemental role in aligning unpaired data, as it provides additional semantic knowledge that enhances the alignment process.

The key challenges here are, first, linking the unpaired data in a meaningful way, specifically by constructing pseudo-pairs between $\mathcal{D}_{\text{ego}}$ and $\mathcal{D}_{\text{exo}}$: $\mathcal{D}_{\text{pseudo-pair}} = \{(v_{\text{ego}}, v_{\text{exo}}) \mid v_{\text{ego}} \in \mathcal{D}_{\text{ego}}, v_{\text{exo}} \in \mathcal{D}_{\text{exo}}\}$ and second, effectively aligning unpaired ego and exo representations in the feature space. Since the pseudo-pairs are not as precisely matched as the data in $\mathcal{D}_{\text{pair}}$, aligning these representations in the feature space requires careful handling of the similarity measure to account for the inherent noise and variability in the unpaired pseudo-paired data.

3.2. Training a Rosetta Stone Translator

Our core insight is to utilize a moderate amount of time-synchronized, paired ego-exo video as a "Rosetta Stone" to facilitate the interpretation of complex associations within unpaired ego and exo data. To achieve this, we train a

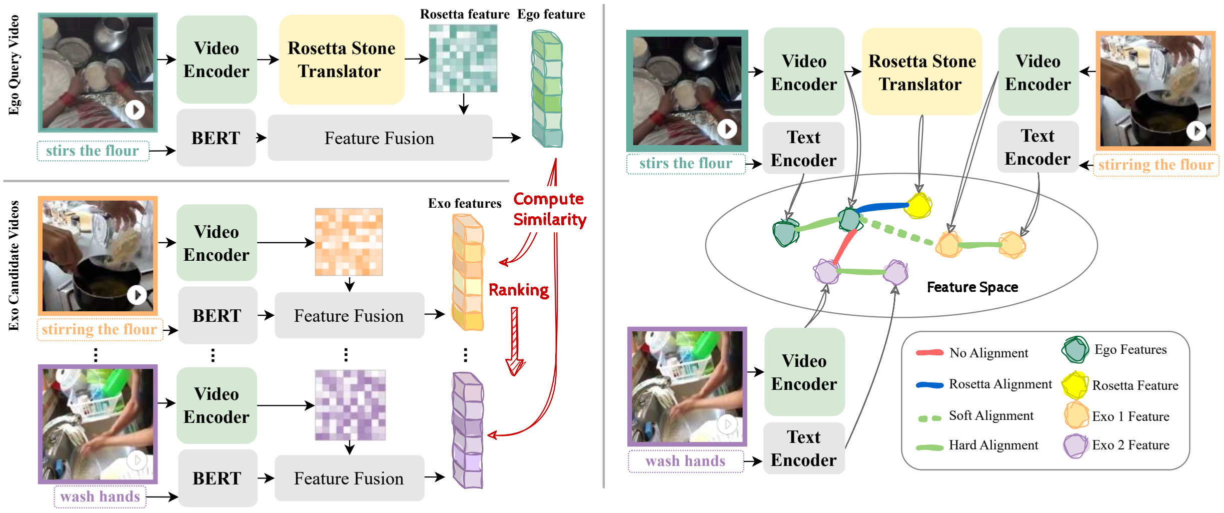

Figure 3. Framework Overview. Left: Our ViewpointRosetta model acts as a bridge to align unpaired ego and exo videos. From an ego query video, RST generates a corresponding exo feature. This hallucinated exo feature is then concatenated with narration embeddings to retrieve the closest match from exo candidate videos. Right: we propose soft view-invariant representation learning. Different from traditional video-text and video-video contrastive learning, our approach involves: (1) assigning weights to pseudo-aligned ego-exo pairs, with higher weights given to pairs showing greater semantic similarity (indicated by the dashed line); (2) RST-synthesized exo feature and the anchor ego feature as the positive pair, to enhance feature alignment across views (highlighted by the blue line).

RST (Rosetta Stone Translator) $\mathcal{T}$, designed as a diffusion model [30] that learns to map between ego and exo viewpoints in the feature space. To train the RST, each video v in the paired dataset $\mathcal{D}_{\text{pair}}$ is first mapped into a latent space with a video encoder f_ϕ pretrained with video-language contrastive learning [51]. This yields meaningful representations from both ego and exo views, resulting in latent features that capture the video's semantic content as well as the inherent discrepancy between viewpoints. Specifically, for each video pair $(v_{\text{ego}}^{(i)}, v_{\text{exo}}^{(i)})$, we obtain the encoded latent features as: $z_{\text{ego}}^{(i)} = f_\phi(v_{\text{ego}}^{(i)}), \quad z_{\text{exo}}^{(i)} = f_\phi(v_{\text{exo}}^{(i)})$.

To perform the translation from z_{ego} to z_{exo}, we adopt the formulation of a denoising diffusion probabilistic model (DDPM) [20]. In the forward diffusion process, we incrementally add Gaussian noise to the exo feature z_{exo} over a fixed number of timesteps. Let $q(z_t|z_{t-1})$ denote the forward process, where t indexes the diffusion timesteps. The forward process is defined as:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{\alpha_t} z_{t-1}, (1-\alpha_t)\mathbf{I}),$$

where α_t controls the variance at each timestep t. Starting from z_{exo}, this process generates a series of noisy latent variables $\{z_t^{(i)}\}_{t=1}^{T}$, where T is the total number of diffusion steps. The reverse process is parameterized by a transformer model ϵ_θ that aims to recover z_{exo} by iteratively denoising z_t given the ego feature z_{ego} as a condition. The reverse process for each timestep t is defined as:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \sigma_t^2\mathbf{I}),$$

where $\mu_\theta(z_t, t)$ is the predicted mean that depends on both z_t and t, and σ_t is the variance term for timestep t.

The training objective for the diffusion-based translator is to minimize the difference between the predicted exo feature and the true exo feature. In practice, the model is trained to minimize the mean squared error (MSE) between the predicted noise and the actual noise added in the forward process. For each pair $(z_{\text{ego}}, z_{\text{exo}})$, the objective is:

$$\mathcal{L}_{\text{RST}} = \mathbb{E}_{z_{\text{ego}}, z_{\text{exo}}, \epsilon\sim\mathcal{N}(0,\mathbf{I}), t}\left[\left\|\epsilon - \epsilon_\theta(z_t, t, z_{\text{ego}})\right\|^2\right]$$

where ϵ is the Gaussian noise added in the forward process.

Through this training process, the RST learns to map ego features z_{ego} into a denoised representation that approximates z_{exo}. We describe how RST is applied to construct pseudo paired ego-exo data and to enhance contrastive learning as data augmentation for the dual encoder in Sections 3.3 and 3.4, respectively. To avoid conflicting learning signals—since the RST needs to translate features between the ego and exo domains, whereas the dual encoder needs to unify them—during fine-tuning of the dual-encoder we keep the RST frozen.

3.3. Matching Unpaired Ego-Exo Video Data

Figure 3 depicts how we create an alignment (matching) mechanism that links each query ego video to a relevant exo video from a pool of candidate exo videos. The high-level idea is to use the pretrained RST as a bridge, mining pseudo-pairs with similar semantics and visual context.

Specifically, for each query ego video $v_{\text{ego}} \in \mathcal{D}_{\text{ego}}$, we first extract its visual feature, using the pretrained video encoder f_ϕ, mapping it into a latent space $z_{\text{ego}} = f_\phi(v_{\text{ego}})$. Then the extracted ego feature z_{ego} is passed through the Rosetta Stone Translator $\mathcal{T}$, to generate a corresponding exo representation $\hat{z}_{\text{exo}} = \mathcal{T}(z_{\text{ego}})$ — hallucinating what would the ego feature look like from the exo perspective.

To encourage a holistic understanding of the input ego video, we further adopt an encoder-based language model BERT [11] to extract semantic information from its text narration. The narration sentence n_{ego} is processed by a BERT model f_{BERT}, which generates a text embedding e_{ego}. To form the final query ego feature h_{ego}, the exo-like feature $\hat{z}_{\text{exo}}$ from the RST and the text embedding e_{ego} are concatenated: $h_{\text{ego}} = \text{concat}(\hat{z}_{\text{exo}}, e_{\text{ego}})$.

For each candidate exo video $v_{\text{exo}}^{(j)} \in \mathcal{D}_{\text{exo}}$, we similarly encode the raw video with f_ϕ and narration $n_{\text{exo}}^{(j)}$ with f_{BERT}, to get the final representation $h_{\text{exo}}^{(j)} = \text{concat}(z_{\text{exo}}^{(j)}, e_{\text{exo}}^{(j)})$.

To determine the best match for the query ego video, we compute the cosine similarity between the query ego feature h_{ego} and each candidate exo feature $h_{\text{exo}}^{(j)}$. The candidate exo videos are ranked based on the similarity scores, and the one with the highest score is selected to be linked.

3.4. Contrastive Learning with Dual-Encoders

As illustrated in Figure 3, we adopt a dual-encoder framework for learning view-invariant representations across the ego and exo perspectives. This framework consists of two components: a video encoder f_θ, which processes a video clip v and maps it to a latent feature representation $f_\theta(v) \in \mathbb{R}^d$, where d is the dimension of the shared feature space, and a text encoder g_ψ, which processes the narration n associated with the video and maps it to the same feature space. This setup ensures that both video and text representations are embedded within a shared latent space, facilitating cross-modal comparisons. f_θ and g_ψ share weights for ego and exo input.

To address the challenge of learning robust view-invariant representations, we leverage a combination of paired data, pseudo-aligned unpaired data, and synthetic data generated by the RST, to bridge the gap between the ego and exo perspectives. Our approach introduces two concepts of alignment — hard alignment and soft alignment in the feature space, enabling flexible and scalable learning across both paired and unpaired data.

The key insight is to handle uncertainty in the alignment of unpaired data. For pseudo-paired data, corresponding views may not always share perfectly matched semantic meanings. To address this challenge, we apply soft alignment, which introduces a weighting factor based on the textual similarity between the ego and exo narrations. This weighting modulates the contrastive loss, with pairs exhibiting higher textual similarity receiving greater weights.

We now describe the formulation of the contrastive loss with soft alignment. For a pseudo-paired sample consisting of an ego video v_{ego} and an exo video v_{exo}, along with their corresponding narrations n_{ego} and n_{exo}, the soft contrastive loss is computed as:

$$\mathcal{L}_{\text{v-v}} = \text{sim}(f_{\text{BERT}}(n_{\text{ego}}), f_{\text{BERT}}(n_{\text{exo}}))\times \left[-\log\left(\frac{\exp(\text{sim}(f_\theta(v_{\text{ego}}), f_\theta(v_{\text{exo}}))/\tau)}{\sum_{x\in\mathcal{B}}\exp(\text{sim}(f_\theta(v_{\text{ego}}), f_\theta(x))/\tau)}\right)\right].$$

The term $\text{sim}(f_{\text{BERT}}(n_{\text{ego}}), f_{\text{BERT}}(n_{\text{exo}}))$ modulates the weight of the contrastive loss based on the textual similarity between ego and exo narration pairs. For naturally paired (synchronized) ego-exo video pairs, this textual similarity term is set to be 1, reflecting perfect alignment.

Additionally, we apply a contrastive loss between each video features and the corresponding text features. By aligning video feature with their respective text descriptions, the model implicitly encourages ego and exo video features to move closer to each other. This happens because the text features serve as a shared anchor — text features with the same semantics are more likely to be assigned to the same subspace. The objective is formulated as:

$$\mathcal{L}_{\text{v-t}} = -\log\left(\frac{\exp(\text{sim}(f_\theta(v_{\text{ego}}), g_\psi(n_{\text{ego}}))/\tau)}{\sum_{x\in\mathcal{B}}\exp(\text{sim}(f_\theta(v_{\text{ego}}), g_\psi(x))/\tau)}\right).$$

To enhance contrastive learning further and provide a reference for the video encoder about what the ego feature would look like in the exo view, we introduce synthetic features generated by the RST feature translator. This augmentation helps bridge the gap between the two perspectives and improves ego-exo alignment. The corresponding contrastive loss is:

$$\mathcal{L}_{\text{v-s}} = -\log\left(\frac{\exp(\text{sim}(f_\theta(v_{\text{ego}}), \mathcal{T}(f_\theta(v_{\text{ego}})))/\tau)}{\sum_{x\in\mathcal{B}}\exp(\text{sim}(f_\theta(v_{\text{ego}}), \mathcal{T}(f_\theta(x)))/\tau)}\right).$$

The final training objective for paired sample is computed by combining the video-video alignment loss and the video-text alignment loss. For pseudo-paired samples, the final training objective is the sum of the video-video, video-text, and video-synthetic data losses. λ_1 and λ_2 are hyperparameters that control the relative importance of each term.

$$\mathcal{L}_{\text{paired}} = \lambda_1\mathcal{L}_{\text{v-v}} + (1-\lambda_1)\mathcal{L}_{\text{v-t}}$$

$$\mathcal{L}_{\text{pseudo-paired}} = \lambda_1\mathcal{L}_{\text{v-v}} + (1-\lambda_1)\mathcal{L}_{\text{v-t}} + \lambda_2\mathcal{L}_{\text{v-s}}.$$

Our ViewpointRosetta approach combines soft alignment for pseudo-paired data with synthetic data generation through the RST, enhancing the flexibility and scalability of contrastive learning across multiple views.

4. Experiments

We validate our approach with multiple cross-view video understanding tasks.

Category	Method	Cross-view Retrieval (HR@5)			Cross-view Recognition		Cross-view Skill Assess. (acc.)
		ego → exo	exo → ego	avg.	top-1 acc.	top-5 acc.	
Egocentric Video Representation Learning	TimeSformer [6]	6.68	6.95	6.82	5.18	14.14	51.58
	EgoVLP [24]	29.33	13.47	21.40	20.33	46.32	54.37
	LaViLa [51]	34.91	12.02	23.47	26.43	55.01	54.10
	LaViLa* [51]	43.63	21.93	32.78	25.51	54.53	54.37
View-invariant Representation Learning	Random Align *	23.33	20.08	21.71	12.19	31.23	52.75
	ActorObserverNet † [37]	29.50	24.85	27.18	15.70	38.45	54.10
	VI Encoder † [17]	29.53	24.40	26.97	14.85	35.39	53.83
	EgoInstructor * [46]	46.04	31.68	38.86	24.15	51.40	54.73
	SUM-L * [42]	47.14	32.77	39.96	24.83	52.08	55.10
	ViewpointRosetta (Ours)	**58.14**	**47.21**	**52.68**	**34.47**	**64.85**	**55.82**

Table 1. Downstream evaluation on three cross-view understanding tasks. * means having access to all the same paired and unpaired data as ours. † means only training with paired data. ViewpointRosetta markedly outperforms all baseline models, demonstrating consistent performance gains across three tasks, highlighting the effectiveness of our approach. All models listed in the table are based on the TimeSformer-B [6] backbone, for the fairest comparison.

4.1. A Benchmark for Cross-View Understanding

To evaluate the effectiveness of view-invariant representations, we introduce a new cross-view understanding benchmark based on the Ego-Exo4D dataset [17], the largest and most comprehensive public ego-exo dataset to date. This dataset offers large-scale perfectly synchronized ego and exo video pairs across a diverse range of domains, making it an ideal testbed for assessing cross-view understanding. This benchmark aims to address two main questions: (1) How view-invariant are the pretrained representations? and (2) How effectively do view-invariant representations facilitate knowledge transfer from exo to ego views?

To address the first question, we define a **cross-view action retrieval** task. Given a query video from either the exo or ego perspective, the objective is to retrieve videos from the opposite view that depict the same action. This task is evaluated using HR@k (Hit Rate@k), which measures the frequency with which the correct match appears within the top k retrieved results. The dataset for this task consists of 2,936 ego and 5,872 exo videos, spanning 188 action classes, allowing evaluations for both exo→ego and ego→exo retrieval performance. We sample the data from the cooking videos from Ego-Exo4D's keystep recognition benchmark. Average video length is 11 seconds.

To explore the second question, we examine how well view-invariant representations support knowledge transfer in two scenarios: action recognition from exo to ego, and skill assessment from exo to ego. This focus on exo-to-ego transfer reflects the typical availability of richer resources and task-specific annotations in the exo domain compared to the ego domain. For **cross-view action recognition**, we frame the task as a 188-way keystep recognition problem. During training, the pretrained representation model is fine-tuned on exo-only data, using 18,518 videos in the training set. At test time, the model is evaluated on ego-only data in a zero-shot setting with a test set of 2,936 videos. This setup assesses the model's ability to generalize its learned representations from exo to ego without additional ego-specific training. In **cross-view skill assessment**, we treat the task as a binary classification problem, where the goal is to classify the skill level displayed in a video as "good" or "bad". During training, the model is fine-tuned on 20,848 exo-only videos and tested in a zero-shot setting on 1,109 ego videos.

Action labels for cross-view retrieval and action recognition tasks are sourced from Ego-Exo4D's keystep annotations, while skill labels for cross-view skill assessment are obtained from the proficiency estimation annotations in Ego-Exo4D. This benchmark provides a comprehensive evaluation of the robustness and transferability of view-invariant representations across retrieval, recognition, and skill assessment tasks in cross-view settings. Our unpaired training data comes from YouTube how-to videos [18, 28] and unscripted egocentric video [16], detailed below.

Note that the original Ego-Exo4D benchmarks are specifically designed around downstream tasks from ego views and do not address cross-view scenarios, which are equally significant in practical applications. Our benchmark supplements the original ones by emphasizing cross-view understanding, sourcing action and skill labels from Ego-Exo4D's comprehensive annotations. We hope this addition serves as a valuable reference for future research in cross-view video understanding.

4.2. Experiment Setup

Baselines We consider two families of models, incorporating a total of seven baselines. For view-invariant learning, we evaluate the following methods: Random Align, which randomly aligns unpaired ego-exo videos; ActorObserverNet [37], which aligns paired data using a triplet loss; the VI encoder from the keystep recognition benchmark of Ego-Exo4D [17], which aligns paired data with an InfoNCE [29] loss; and SUM-L [42] and EgoInstructor [46], which align unpaired ego-exo videos based on language semantics. For general video representation learning baselines, we consider TimeSformer, which uses spatial attention modules initialized with weights from CLIP [33], and video-language pretraining models such as EgoVLP [24] and LaViLa [51]. ActorObserverNet [37] and the VI encoder [17] are trained exclusively on paired data only, due to their design.

Implementation Details In our dual encoder architecture, the video encoder is a TimeSformer [6] with spatial attention modules initialized from a ViT [12] model that was contrastively pre-trained on large-scale paired image-text data as in CLIP [33]. We sample 4 frames per clip during fine-tuning on downstream tasks. For EgoVLP [24], we use 16 frames per clip due to the fixed sampling rate in the public checkpoint. The text encoder follows a 12-layer Transformer architecture as in [51]. For the Rosetta Stone Translator, we use the DiT [30] model architecture. For the BERT model, we adopt Sentence Transformers [34]. Pretraining is conducted over 5 epochs, while fine-tuning on downstream tasks is performed over 200 epochs. Extra details can be found in the supplementary.

Pretraining Datasets For paired data, we use the Ego-Exo4D dataset [17], which contains approximately 850k time-synchronized, multi-view ego-exo video pairs. For unpaired ego data, we utilize Ego4D [16], the largest egocentric video dataset available. Each clip's interval is determined by the pairing strategy outlined in [24], resulting in a pool of 562k ego video clips paired with human narrations. The average clip length is 1 second. For unpaired exo data, we use the HTM-AA dataset [18], a temporally aligned version of HowTo100M [28]. Our training set contains approximately 1.25M exo video-narration pairs.

Due to computational constraints, our pretraining setup focuses on cooking samples, yet the dataset remains vast — over 2.5M clips. Moreover, cooking videos in Ego-Exo4D offer high diversity, spanning 188 action classes and 60 distinct environments.

4.3. Main Results

Overall Performance Looking at Table 1, our ViewpointRosetta outperforms all other models across all three benchmark tasks. The cross-view retrieval task serves as a direct measure of the view-invariance of the pretrained representations. Specifically, comparing with ActorObserverNet [37] and VI Encoder [17], which rely exclusively on paired data, ViewpointRosetta's use of both paired and unpaired data significantly enhances its view-invariance. SUM-L [42], which aligns unpaired videos based on language semantics, performs better than other view-invariant baselines but still falls short of ViewpointRosetta. Importantly, this shows that our use of the Rosetta Stone Translator for feature alignment is more effective than language-based alignment alone in capturing nuanced view-invariant representations. Also, results in cross-view recognition and skill assessment tasks reveal that ViewpointRosetta enables more effective knowledge transfer from exo to ego than existing approaches. Our margins to the next best baseline are

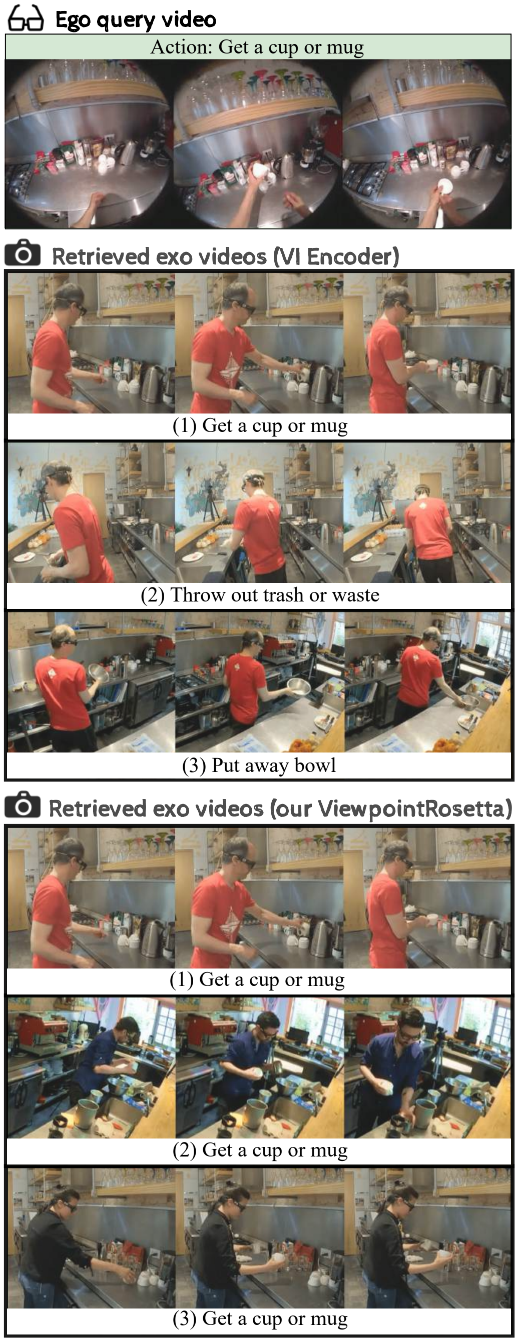

Figure 4. Compared to the VI Encoder [17], our ViewpointRosetta unlocks unpaired data to capture rich semantic information, enabling retrieval of samples that are not only visually similar to the input view but also semantically related.

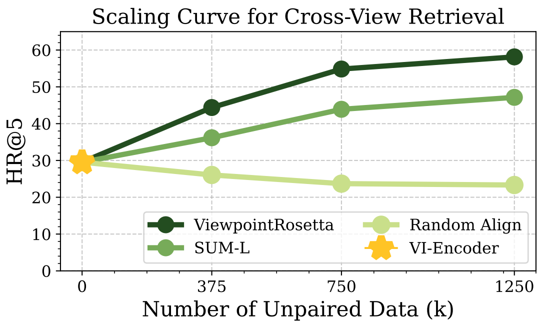

Figure 5. ViewpointRosetta is better at "unlocking" and delivers better view-invariant features when the data volume increases.

generally large. The smallest margins are for skill assessment, which we attribute to the inherent difficulty of the task; most methods perform only a bit better than random chance.

Note that all models in Table 1 use the TimeSformer-B backbone pretrained on CLIP data as the backbone, allowing a fair apples-to-apples comparison. Preliminary experiments indicate that replacing the backbone with EgoVideo [31] only improves the absolute performance of ViewpointRosetta (+1.2% on the cross-view retrieval task).

Overall, the results show that ViewpointRosetta not only learns robust view-invariant representations but also supports practical cross-view applications by facilitating effective knowledge transfer from exo to ego.

Scaling behavior Figure 5 illustrates how different methods of aligning unpaired data (Random Align vs. SUM-L [42] vs. ours) perform on cross-view retrieval as the volume of unpaired data gradually increases from 0 to 1.25M.

Note that VI-Encoder [17] uses no unpaired data due to its design limit, and hence is a single point on the left side. ViewpointRosetta is significantly more effective in leveraging unpaired data for view-invariant representation learning compared to other methods. By capitalizing on unpaired data to reinforce conceptual understanding (rather than exact matches), our approach scales not just in data volume but in depth of comprehension. Our model's RST provides a direct, feature-level mapping between views that captures detailed visual correspondences, allowing it to unlock the potential of unpaired data more effectively. This explicit cross-view translation enables ViewpointRosetta to create synthetic feature pairs, build a more structured representation space, and handle viewpoint-specific details, all of which are essential for effective scaling.

Retrieval Qualitative Examples In Figure 4, we present a qualitative example of cross-view retrieval. We observe that the VI Encoder [17], which is trained only on paired data, tends to retrieve exo videos that are visually similar to the input but often lack semantic alignment with the action depicted. In contrast, by leveraging unpaired data, our model goes beyond surface-level visual similarities and retrieves results with meaningful action-based alignment.

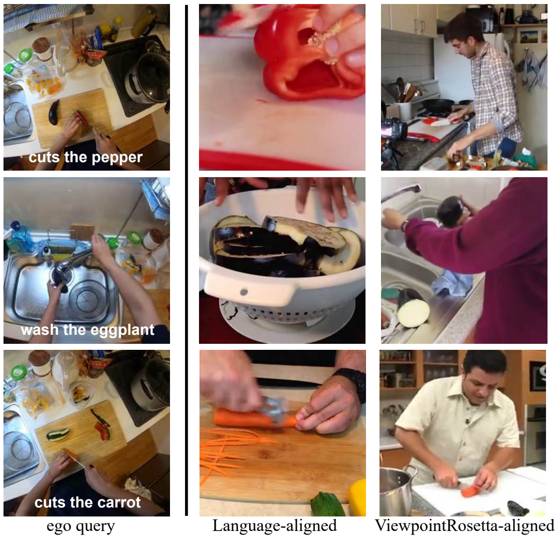

Figure 6. ViewpointRosetta's RST tends to align videos which are more visually similar, contain more environment info, and feature larger view point change.

Behavior of RST Figure 6 illustrates the top-1 alignment choices using language guidance compared to our RST approach. Language-based alignment tends to favor close-up shots, while the RST aligns based on both semantic and visual similarities, producing matches that more closely resemble the original action as seen from the exo view. This often includes wide-angle shots that capture broader environmental cues and scene layout, resulting in alignments that preserve both the context and spatial relationships present in the egocentric view.

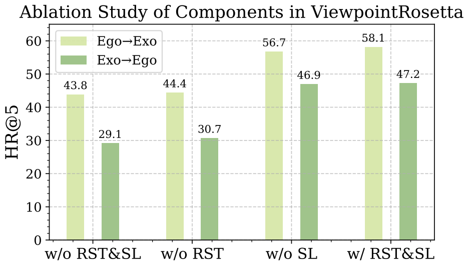

Figure 7. Ablation of components of our ViewpointRosetta.

Ablation Study The ablation study in Figure 7 provides insights into the individual and combined contributions of Viewpoint Rosetta Stone (RST) component and Contrastive learning with Soft Alignment (SL) in our framework, measured in terms of HR@5 scores for cross-view ego-exo retrieval. It reveals that the RST Feature Translator is fundamental to achieving strong view-invariance in our ViewpointRosetta, while Soft Alignment further enhances performance by handling imperfect matches in unpaired data. The best results come from combining both. The significant drop in performance when removing RST demonstrates that our RST is crucial for achieving the best results, as traditional language-based alignment alone is insufficient.

5. Conclusion

We propose a novel approach for learning ego-exo view-invariant video representations by leveraging both paired and unpaired data. Central to our method is a diffusion-based Rosetta Stone Translator (RST), trained on synchronized ego-exo videos, which deciphers complex relationships within unpaired data and generates pseudo ego-exo pairs to enable multi-view contrastive learning. Additionally, we introduce a new cross-view understanding benchmark derived from the Ego-Exo4D dataset, setting the stage for future advancements in ego-exo cross-view understanding. Our work has significant implications for applications such as robot learning and human skill acquisition in augmented reality (AR), where an egocentric actor must interpret—or even replicate—the actions of a demonstrator observed from an exocentric perspective.

Acknowledgment UT Austin is supported by NSF Grants 2019844, 2112471, AF 1901292, CNS 2148141, Tripods CCF 1934932, Tripods 2217069, NSF AI Institute for Foundations of Machine Learning (IFML) 2019844, the Texas Advanced Computing Center (TACC) and research gifts by Western Digital, Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program, UT Austin Machine Learning Lab (MLL), Cisco and the Stanly P. Finch Centennial Professorship in Engineering.

References

[1] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 253–268. Springer, 2016. 2

[2] Shervin Ardeshir and Ali Borji. Egocentric meets top-view. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1353–1366, 2018. 2

[3] Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 171:61–68, 2018. 2

[4] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23066–23078, 2023. 2

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3

[6] 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. 6, 7

[7] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024. 3

[8] Feng Cheng, Mi Luo, Huiyu Wang, Alex Dimakis, Lorenzo Torresani, Gedas Bertasius, and Kristen Grauman. 4diff: 3d-aware diffusion model for third-to-first viewpoint translation. In *ECCV*, 2024. 2

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 1

[11] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[12] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[13] 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. *arXiv e-prints*, pages arXiv–2408, 2024. 2

[14] Mohamed Elfeki, Krishna Regmi, Shervin Ardeshir, and Ali Borji. From third person to first person: Dataset and baselines for synthesis and retrieval. *arXiv preprint arXiv:1812.00104*, 2018. 2

[15] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5125–5133, 2017. 2

[16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 6, 7

[17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, 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*, pages 19383–19400, 2024. 1, 2, 3, 6, 7, 8

[18] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2906–2916, 2022. 6, 7

[19] Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons' points of view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–85, 2018. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4

[21] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. 2

[22] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023. 3

[23] 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*, pages 6943–6953, 2021. 2

[24] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 2, 6, 7

[25] Gaowen Liu, Hao Tang, Hugo Latapie, and Yan Yan. Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1843–1847. IEEE, 2020. 2

[26] Hongchen Luo, Kai Zhu, Wei Zhai, and Yang Cao. Intention-driven ego-to-exo video generation. *arXiv preprint arXiv:2403.09194*, 2024.

[27] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In *European Conference on Computer Vision*, pages 407–425. Springer, 2025. 2

[28] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 2, 6, 7

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 7

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 2, 4, 7

[31] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, and Yu Qiao. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024. 8

[32] 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*, pages 5285–5297, 2023. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[34] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 7

[35] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 2

[36] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. *Proceedings of International Conference in Robotics and Automation (ICRA)*, 2018. 2

[37] 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*, pages 7396–7404, 2018. 2, 3, 6, 7

[38] 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. 2

[39] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3224–3234, 2021. 2

[40] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5250–5261, 2023. 2

[41] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13157–13166, 2022. 2

[42] 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*, pages 3307–3317, 2023. 2, 6, 7, 8

[43] Yangming Wen, Krishna Kumar Singh, Markham Anderson, Wei-Pang Jan, and Yong Jae Lee. Seeing the unseen: Predicting the first-person camera wearer's location and pose in third-person scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3446–3455, 2021. 2

[44] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15802–15812, 2023. 3

[45] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. *ArXiv*, abs/2302.00402, 2023. 3

[46] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. *arXiv preprint arXiv:2401.00789*, 2024. 6, 7

[47] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–652, 2018. 2

[48] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710, 2023. 2

[49] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. First-and third-person video co-analysis by learning spatial-temporal joint attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[50] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 3

[51] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 2, 4, 6, 7