# Sound Bridge: Associating Egocentric and Exocentric Videos via Audio Cues

Sihong Huang[1,5,*], Jiaxin Wu[2,*], Xiaoyong Wei[3, 2, 5,†], Yi Cai[1,†], Dongmei Jiang[5], Yaowei Wang[4,5]
[1]South China University of Technology, [2]The Hong Kong Polytechnic University
[3]Sichuan University, [4]Harbin Institute of Technology, Shenzhen, [5]Peng Cheng Laboratory
{fthuangsh, ycai}@mail.scut.edu.cn, jiaxwu@polyu.edu.hk, cswei@scu.edu.cn
jiangdm@pcl.ac.cn, wangyaowei@hit.edu.cn

## Abstract

*Understanding human behavior and environmental information in egocentric videos is very challenging due to the invisibility of some actions (e.g., laughing and sneezing) and the local nature of the first-person view. Leveraging the corresponding exocentric video to provide global context has shown promising results. However, existing visual-to-visual and visual-to-textual Ego-Exo video alignment methods struggle with the issue that some activities may have non-visual overlap. To address this, we propose using sound as a bridge, as audio is often consistent across Ego-Exo videos. However, direct audio-to-audio alignment lacks context. Thus, we introduce two context-aware sound modules: one aligns audio with vision via a visual-audio cross-attention module, and another aligns text with sound closed caption generated by LLM. Experimental results on two Ego-Exo video association benchmarks show that each of the proposed modules enhances the state-of-the-art methods. Moreover, the proposed sound-aware egocentric or exocentric representation boosts the performance of downstream tasks, such as action recognition of exocentric videos and scene recognition of egocentric videos. The code and models can be accessed at* https://github.com/shhuangcoder/SoundBridge.

## 1. Introduction

Egocentric video analysis has been receiving extensive attention recently [10, 12, 34] due to its potential applications in robotic vision analysis, embedded AI, etc. Compared to the large-scale and in-depth studies on exocentric video analysis tasks, research on egocentric videos is still underexplored. Moreover, there are significant differences between egocentric and exocentric videos; thus, many findings and methodologies derived from an exocentric perspective are not directly applicable to egocentric video analysis.

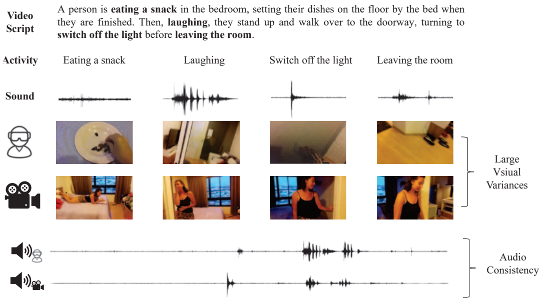

Video Script: A person is **eating a snack** in the bedroom, setting their dishes on the floor by the bed when they are finished. Then, **laughing**, they stand up and walk over to the doorway, turning to **switch off the light** before **leaving the room**.

Figure 1. Given a video shooting script, egocentric and exocentric views can exhibit significant visual variances for the same activity, while their audio signals remain relatively consistent.

The main difference between egocentric and exocentric videos is that they capture events from different perspectives. First-person view videos focus on local and detailed information about the current activity, whereas third-person view videos emphasize global and environmental context. An example is shown in Figure 1. Although egocentric and exocentric videos follow the same video shooting script, they exhibit significant visual variations. For example, in the case of the same activity, such as eating a snack, the egocentric view primarily captures hand-object interactions, such as a hand picking up a raisin, while the location remains unknown. In contrast, the exocentric view captures the overall scene of the event but lacks finer details—for example, it may show a person lying on a bed eating food, but the specific type of food remains unclear. Moreover, certain activities, such as laughing and sneezing, do not produce distinct visual cues in egocentric videos apart from camera shaking. Consequently, when attempting to align visual representations from exocentric to egocentric videos, there is often no overlapping visual content to match.

However, recent studies [12, 34] have shown that exocentric videos can supplement the understanding of their semantically aligned egocentric counterparts. In this paper, we also focus on constructing an Ego-Exo alignment space to facilitate egocentric video analysis. Previous studies on

[*] Equal Contribution. † Corresponding Authors.

egocentric and exocentric video alignment have primarily focused on matching visual content [35] or aligning both video types to an anchor (i.e., their semantic descriptions or a shared textual representation) [12, 34]. However, these alignment strategies may be insufficient due to the substantial visual variations between views. In contrast to the significant differences in view-dependent visual content, we observe that audio signals for the same event remain consistent across views. For example, in the case of laughing, the laughter sound—whether captured in an egocentric or exocentric video—can infer that someone is laughing.

In this paper, we propose using audio signals as a bridge to model the egocentric-exocentric video alignment space. However, directly linking audio signals across videos without context may be suboptimal, as audio information alone is not sufficiently discriminative [2]. For example, the sound of a click associated with the action "switch off the light" in Figure 1 could also occur in other activities, such as "flipping a power switch on a household appliance" or "turning off a desk lamp". To address this limitation, we propose constructing contextual audio features for alignment by incorporating visual and textual context. An overview of our proposed method is presented in Figure 2. Specifically, we introduce sound-text and sound-vision modules to bridge Ego-Exo videos. The sound-vision module follows existing methods [13, 18] to build contextual audiovisual representations through cross-attention. Meanwhile, the sound-text module utilizes a large language model (LLM) to generate contextual audio descriptions and construct sound-aware textual representations. In particular, we provide the LLM with semantic text describing the video scenario and prompt it to generate an audio description of the actions occurring within the scene in a closed-caption format [15]. Additionally, to ensure that the generated sound captions emphasize unique information within the paired cross-view videos rather than commonly occurring content in the dataset, we filter out frequently appearing and less informative concepts (e.g., person) from the generated descriptions. As a result, the output sound caption for the example in Figure 1 is "[Eating the snack] (munching), [dishes on the floor] (crashing and scraping), [Laughing] (giggles), [leaving the room] (Footsteps), [turns off the light] (click)". The generated sound caption and the semantic text are subsequently processed through a cross-attention block to learn a sound-aware textual representation. Finally, we treat the textual representation as an anchor and align the audiovisual representations of egocentric and exocentric videos to it, thereby learning an Ego-Exo semantic feature space. We conduct extensive experiments on two Ego-Exo video retrieval benchmarks, namely EgoExoLearn [12] and CharadesEgo [1]. The results demonstrate that the learned sound-aware features outperform state-of-the-art approaches in both egocentric-exocentric video retrieval and text-to-egocentric/exocentric video retrieval. Moreover, the learned Ego- and Exo- features enhance the performance of downstream tasks, such as exocentric video action recognition and egocentric video scene recognition.

## 2. Related Work

### 2.1. Egocentric Video Analysis

The number of research studies on egocentric video analysis is growing rapidly, as this task is a fundamental component of many applications [10, 12], such as robotic vision analysis and embedded AI. However, egocentric videos are fundamentally different from traditional exocentric videos. Egocentric videos capture a unique immersive viewpoint from a first-person perspective, directly reflecting an individual's interaction with the current activity, whereas exocentric videos provide an external view of human interactions within an environment. This distinct perspective in egocentric videos introduces a range of challenges for human activity analysis, including hand detection [5, 9, 26], action recognition [8, 22, 28], human-computer interaction [16, 21, 33], and applications in virtual and augmented reality [14, 19, 29]. Despite these differences, recent studies [12, 34] have demonstrated that egocentric video analysis can benefit from the contextual information provided by corresponding exocentric videos. For example, Xu et al. [34] show that related exocentric videos can enhance egocentric video captioning. Similarly, Huang et al. [12] demonstrate that the exocentric perspective provides complementary information that aids robots in performing complex tasks. In this paper, we also focus on developing effective Ego-Exo video alignment techniques to enhance egocentric video analysis.

### 2.2. Ego-Exo Video Representation Learning

The goal of Ego-Exo video representation learning is to construct a semantic space in which egocentric and exocentric videos with similar semantics are positioned closely, while dissimilar Ego-Exo videos are placed farther apart [12]. Early studies on Ego-Exo representation learning align cross-view videos using visual information. For example, AE2 [35] extracts visual regions of hands and active objects and performs temporal alignment on these extracted regions across views, assuming that the visual content of a fine-grained action remains consistent across perspectives. AE2 constructs four action-specific cross-view datasets for evaluation: break egg, pour milk, pour liquid, and tennis forehand. Recent approaches aim to learn a semantic Ego-Exo representation space by aligning videos to either a shared semantic text or separate textual descriptions. For example, EgoExo [12] employs implicit Ego-Exo alignment by associating each video with a view-dependent text, based on the assumption that semantically similar texts will be positioned closely in the learned space. Beyond implicit alignment, EgoExoNCE [34] takes this further by explicitly aligning Ego-Exo videos to a mutual text anchor. This shared text is generated by extracting key nouns and verbs from both the egocentric and exocentric textual descriptions. In this paper, we follow the pipeline of previous work by aligning Ego-Exo videos to a mutual text anchor. However, rather than directly matching entire videos with full-text descriptions without attention, we propose using audio as a bridge to address the challenge of visual variance. Our approach incorporates sound-vision and sound-text modules to bridge the gap between egocentric and exocentric perspectives. To the best of our knowledge, we are the first to integrate sound into Ego-Exo video alignment.

### 2.3. Sound-aware Representation Learning

Numerous studies have demonstrated that audio signals complement visual signals in video analysis tasks [18]. For example, the Eclipse method has shown that audio signals can serve as an alternative to visual signals in long-range video retrieval, significantly accelerating the search process [18]. Consequently, many approaches have sought to develop sound-aware representations. Several works have incorporated visual information as context to create audiovisual features for exocentric video analysis, applying them to tasks such as video retrieval [13, 18], action recognition [20], and video summarization [11]. More recently, audio-based methods have also been explored in egocentric video analysis, such as active speaker spatial detection [25]. Additionally, Ibrahimi et al. [13] integrate textual context to construct audio-textual representations. Common approaches to incorporating audio signals into video analysis include appending audio features to visual or textual features [20, 25] or learning audio-textual/visual cross-attention representations [13, 18]. Another method involves converting audio signals into text using Automatic Speech Recognition (ASR) and subsequently appending the ASR-derived text features to visual features [11]. Our proposed sound-vision model follows prior work by learning contextual audiovisual features through cross-attention. Additionally, we introduce a novel approach to learning audio-textual representations in our sound-text module, which leverages a large language model (LLM) to generate sound captions.

## 3. Sound-aligned Ego-Exo Representation Learning

The architecture of our proposed method is illustrated in Figure 2. The main objective is to construct a sound-aligned Ego-Exo representation space that integrates local (Ego) and global (Exo) visual features along with their corresponding semantic concepts. Specifically, to address the challenge of aligning semantic text with the significant visual variations present in Ego-Exo videos, we leverage sound as a unifying element. This is achieved by incorporating sound-vision and sound-text modules to generate sound-aware representations. Subsequently, we align the sound-aware video features with the sound-aware textual features to learn a unified representation space.

### 3.1. Sound-Vision Module

Given a semantically aligned egocentric and exocentric (ego-exo) video pair $(v^{ego}, v^{exo})$ and their corresponding audio signals $(a^{ego}, a^{exo})$, we propose a sound-vision cross-attention module to learn their sound-aware visual representations. The sound-vision module consists of a video encoder $E_v(x)$, an audio encoder $E_a(x)$, and a multi-head cross-attention block $Multihead(Q, K, V)$ [27]. The module inputs a video $v$ and its corresponding audio signal $a$ and outputs sound-aware visual feature $f_{sv}$ as:

$$f_v = E_v(v), f_a = E_a(a), \tag{1}$$

$$CrossAB(Q, K, V) = Softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \cdot V, \tag{2}$$

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{3}$$

$$f_{sv} = MultiHead(Q(f_a), K(f_v), V(f_v)). \tag{4}$$

where the $Q(x)$, $K(x)$, $V(x)$ map the input feature $x$ to a hidden space though linear projections and $h$ represents the number of head. The egocentric and exocentric sound-vision modules in Figure 2 share the same architecture, as illustrated in Eqs. (1)–(4), but with view-dependent weights. Specifically, we feed egocentric videos and exocentric videos and their audio signals to the sound-vision module to obtain sound-aware representations as:

$$f_v^{ego} = E_v^{ego}(v^{ego}), f_v^{exo} = E_v^{exo}(v^{exo}). \tag{5}$$

$$f_a^{ego} = E_a(a^{ego}), f_a^{exo} = E_a(a^{exo}). \tag{6}$$

$$f_{sv}^{ego} = MultiHead(Q(f_a^{ego}), K(f_v^{ego}), V(f_v^{ego})) \tag{7}$$

$$f_{sv}^{exo} = MultiHead(Q(f_a^{exo}), K(f_v^{exo}), V(f_v^{exo})) \tag{8}$$

### 3.2. Sound-Text Module

Given a text $t$ that captures the semantics of Ego-Exo videos, we use an LLM to generate a sound caption $c$ and propose a sound-text cross-attention module to learn a sound-aware text representation $f_{st}$. The sound-text module shares the same architecture as the sound-vision module but uses different encoders:

$$f_t = E_t(t), f_c = E_t(c), \tag{9}$$

$$f_{st} = MultiHead(Q(f_c), K(f_t), V(f_t)). \tag{10}$$

Specifically, the sound caption $c$ is generated from the given text $t$ by an LLM. With $t$ providing the video context, we prompt the LLM to generate sound descriptions for all the actions in the videos as: $c = LLM(t, prompt_{Txt2Sound})$. The sound description needs to contain the source of the sound and the sound effect, formatted in Closed Caption (CC) style [15]. Closed captions (CC) are a specialized type of caption designed for people who are deaf or hard of

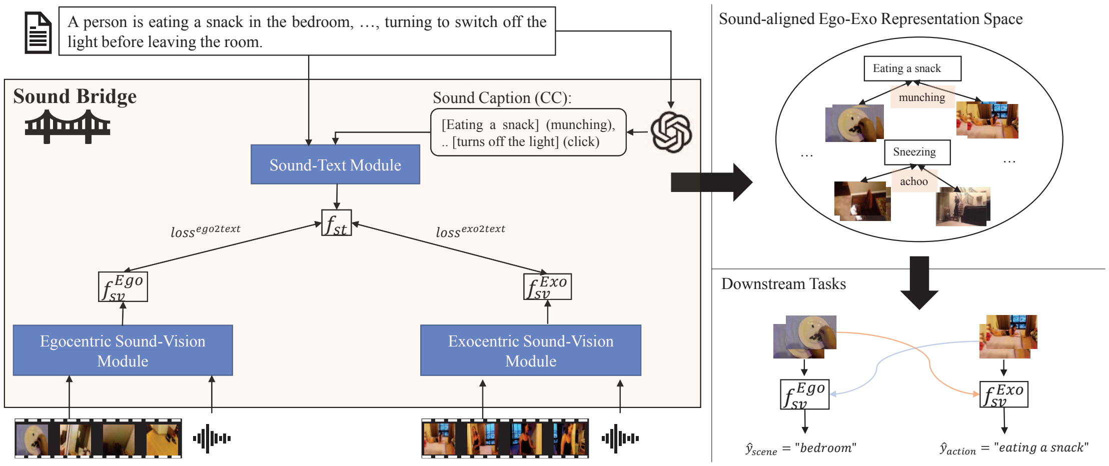

A person is eating a snack in the bedroom, …, turning to switch off the light before leaving the room.
Sound Bridge
Sound Caption (CC):
[Eating a snack] (munching), .. [turns off the light] (click)
Sound-Text Module
$f_{st}$
$loss^{ego2text}$
$loss^{exo2text}$
$f_{sv}^{Ego}$
$f_{sv}^{Exo}$
Egocentric Sound-Vision Module
Exocentric Sound-Vision Module
Sound-aligned Ego-Exo Representation Space
Eating a snack
munching
Sneezing
achoo
…
…
Downstream Tasks
$f_{sv}^{Ego}$
$f_{sv}^{Exo}$
$\hat{y}_{scene}$ = "bedroom"
$\hat{y}_{action}$ = "eating a snack"

Figure 2. Overview of the proposed method. We leverage sound as a bridge to connect Ego-Exo videos with their corresponding semantic text to build a sound-aligned representation space.

hearing. They additionally include an audio description in parentheses and indicate the source of the sound in brackets. For example, for a given text $t$ ="a person turning off the light", the generated sound caption is $c$ ="[a person turning off the light] (click)". Furthermore, to ensure that the sound caption focuses on the unique information of the paired videos, we filter out frequently appearing concepts (e.g., "a person") in CC. To achieve this, we use the LLM to refine $c$ to $\tilde{c}$ by providing the frequently appeared concepts, i.e., $\tilde{c} = LLM(c, frequentConcepts, prompt_{soundRefine})$. Due to space limitations, details of all prompts are provided in the supplementary material. We input the filtered and refined closed sound caption $\tilde{c}$ along with the text $t$ into the sound-text module to obtain the sound-aware text representation: $f_{st} = MultiHead(Q(f_{\tilde{c}}), K(f_t), V(f_t))$.

### 3.3. Sound-aware Representation Space Learning

Given $m$ triplets of sound-aware representations for the Ego-Exo video pairs and their corresponding texts: $\{(f_{sv}^{ego(1)}, f_{sv}^{exo(1)}, f_{st}^{(1)}),\ldots,(f_{sv}^{ego(m)}, f_{sv}^{exo(m)}, f_{st}^{(m)})\}$, the visual representations are trained to be closer to the text representation within the same triplet while remaining distant from other text representations using two contrastive losses.

$$loss^{\text{ego2text}} = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{\exp(\text{sim}(f_{sv}^{ego(i)}, f_{st}^{(i)})/\tau)}{\sum_{k\neq i}\exp(\text{sim}(f_{sv}^{ego(i)}, f_{st}^{(k)})/\tau)} \tag{11}$$

$$loss^{\text{exo2text}} = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{\exp(\text{sim}(f_{sv}^{exo(i)}, f_{st}^{(i)})/\tau)}{\sum_{k\neq i}\exp(\text{sim}(f_{sv}^{exo(i)}, f_{st}^{(k)})/\tau)} \tag{12}$$

where $\tau$ is the temperature parameter. We use the sum of these two losses to train the sound-aligned Ego-Exo representation space.

For the Ego-Exo video retrieval and text-video retrieval tasks, we compute the similarity between $(f_{sv}^{ego}, f_{sv}^{exo})$ and $(f_{sv}, f_{st})$ to obtain ranked lists.

## 4. Sound-aligned Space for Downstream Tasks

To verify the effectiveness of the learned sound-aligned Ego-Exo representation space, we apply the learned sound-aware visual representation to two tasks: scene recognition on egocentric and action recognition on exocentric video. Scene recognition is particularly challenging in an egocentric view because the ego view typically captures local information, such as hand-object interactions. Here, we investigate whether the proposed egocentric sound-vision representation can learn global and environmental information from an exocentric view, thereby enhancing scene recognition performance. We predict the scene classes as follows:

$$\hat{y}_{scene} = softmax(FC(f_{sv}^{ego})). \tag{13}$$

where $FC$ is a fully connected layer that maps the sound-aligned feature to scene classes.

Similarly, we predict the action label using the exocentric sound-aware visual feature as:

$$\hat{y}_{action} = softmax(FC(f_{sv}^{exo})). \tag{14}$$

where $FC$ is a fully connected layer that maps the sound-aligned feature to action classes. We train the FC layers using cross-entropy loss with the ground truth labels.

## 5. Experiments

In this section, we first evaluate the effectiveness of the sound-aware representation space on Ego-Exo video retrieval and text-video retrieval tasks. Next, we apply the sound-aware visual representations to two downstream tasks: scene recognition in egocentric videos and action recognition in exocentric videos. Finally, we conduct ablation studies to verify the effectiveness of our proposed modules.

## 5.1. Experimental Settings

### 5.1.1 Datasets and Evaluation Metric

We conduct the retrieval tasks on two Ego-Exo benchmarks. One benchmark consists of strictly paired Ego-Exo videos (i.e., CharadesEgo [1]), which are recorded using the same shooting script. The other benchmark consists of loosely paired Ego-Exo videos (i.e., EgoExoLearn [12]), where most video pairs share only the same verbs and nouns in their descriptions. CharadesEgo [1] is a large-scale Ego-Exo dataset featuring videos of daily activities. It comprises 7,985 video clips with a total duration of approximately 80 hours. The dataset is divided into training, validation, and test subsets, containing 5,338, 1,334, and 1,313 video pairs, respectively, along with corresponding shooting script descriptions. EgoExoLearn [12] is another large-scale Ego-Exo dataset, consisting of 747 video clips with a total duration of 120 hours. Following [34], we construct 36,372 Ego-Exo pairs by aligning nouns and verbs in the video text descriptions. We use the official data split, which includes 36,373 video pairs for training, 800 for validation, and 2,200 for testing. The downstream tasks are also conducted on CharadesEgo [1] and EgoExoLearn [12]. CharadesEgo contains 16 scene categories (e.g., kitchen and bedroom), while EgoExoLearn covers 19 action categories from both daily life and laboratory scenes (e.g., pick-up and washing). We use CharadesEgo for egocentric video scene recognition and EgoExoLearn for exocentric video action recognition. Following previous works [12, 34], we report top-1 accuracy for both retrieval and recognition tasks.

### 5.1.2 Implementation Details

We use the video transformer from [3] with a ViT-B/16 backbone initialized with CLIP weights [23] as the video encoder $E_v(x)$. For the sound-vision module, we apply the audio encoder from [6] as $E_a(x)$, keeping it frozen during training. In the sound-text module, we use the pre-trained CLIP textual encoder as $E_t(x)$ to encode both semantic text and sound captions. The output dimension of the sound-aligned Ego-Exo representation space is 256. We use Llama3-8B as the LLM to generate sound captions. In the multi-head cross-attention module, we set the number of heads to $h = 4$. The temperature parameter $\tau$ is set to 0.07. We utilize 8 V100 GPUs to train the sound-aware representation space with a learning rate of 3e-5, using AdamW as the optimizer for improved convergence and stability.

Table 1. Retrieval result comparison (top-1 accuracy) of Exo-Ego video alignment task on two benchmarks.

| Method | EgoExoLearn | | | CharadesEgo | | |
|---|---|---|---|---|---|---|
| | Eg2Ex | Ex2Eg | Avg | Eg2Ex | Ex2Eg | Avg |
| FrozenInTime [3] | 14.10 | 13.40 | 13.75 | 6.73 | 4.96 | 5.85 |
| LaViLa [36] | 28.70 | 25.70 | 27.20 | 58.24 | 46.93 | 52.59 |
| EgoVLP [17] | 32.10 | 28.9 | 30.50 | 58.27 | 60.76 | 59.52 |
| InternVideo [32] | 30.60 | 21.70 | 26.15 | 53.55 | 61.23 | 57.39 |
| EgoExo [12] | 49.00 | 45.30 | 47.15 | 68.32 | 62.06 | 65.19 |
| EgoInstructor [34] | 47.10 | 45.82 | 46.46 | 62.06 | 56.38 | 59.22 |
| SoundBridge | **50.27** | **50.73** | **50.50** | **77.66** | **71.63** | **74.65** |

It is worth noting that for video pairs with high synchronization, applying the Dynamic Time Warping (DTW) algorithm [24] to align the two audio sequences before inputting the audio signals into the cross-attention module can help mitigate the impact of varying sound speeds. We use this setting by default for videos in CharadesEgo and provide an ablation study on DTW in the supplementary material.

### 5.1.3 Baseline Models

We include several types of models as baselines. These include recent text-video representation models pre-trained on large-scale exocentric datasets, such as FrozenInTime [3] and InternVideo [32]. Additionally, we include pre-trained models on large-scale egocentric datasets, namely LaViLa [36] and EgoVLP [17]. Furthermore, we compare our method with two recent approaches designed for Ego-Exo video alignment, namely EgoExo [12] and EgoInstructor [34]. For the downstream tasks, we also include task-specific models, namely VideoMAE [30] for exocentric action recognition and InternImage [31] for egocentric scene recognition.

## 5.2. Results on Ego-Exo Video Alignment

Table 1 compares the results of Ego-to-Exo and Exo-to-Ego video retrieval with other baselines. Specifically, we report the zero-shot retrieval performance of the FrozenInTime model and the fine-tuned results of other methods on the EgoExoLearn and CharadesEgo datasets. Our proposed method, SoundBridge, consistently outperforms other baselines, including pre-trained egocentric models, pre-trained exocentric models, and recent Ego-Exo models, on both Ego-to-Exo and Exo-to-Ego video retrieval tasks across the two benchmarks. We are the first model to achieve a top-1 accuracy exceeding 50% on the EgoExoLearn dataset. Specifically, SoundBridge achieves higher retrieval accuracy than EgoExo [12] and EgoInstructor [34] by 8.13% and 15.93% on Ego-to-Exo video retrieval, and by 13.71% and 18.88% on Exo-to-Ego video retrieval, respectively. The performance boost primarily comes from the incorporation of sound-aware modules in the alignment process. For example, given an egocentric video query recording *picking up the garlic with the left hand*, the EgoExo model ranks an exocentric video of *put the ginger down with both hand on the iron bowl* in the top-1 position. In contrast, our

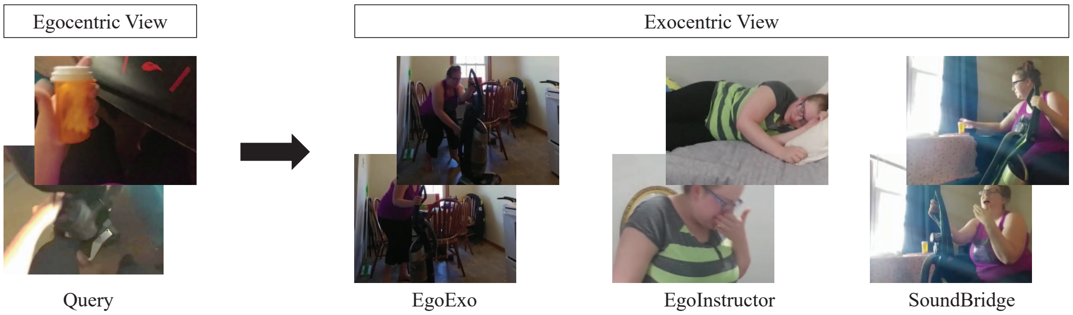


Figure 3. Visualization of an Ego-to-Exo video retrieval example, compared with EgoExo [12] and EgoInstructor [34], on an egocentric video query looking for *A person is working on a vacuum in the man cave, then sneezes and reaches for some medicine* .

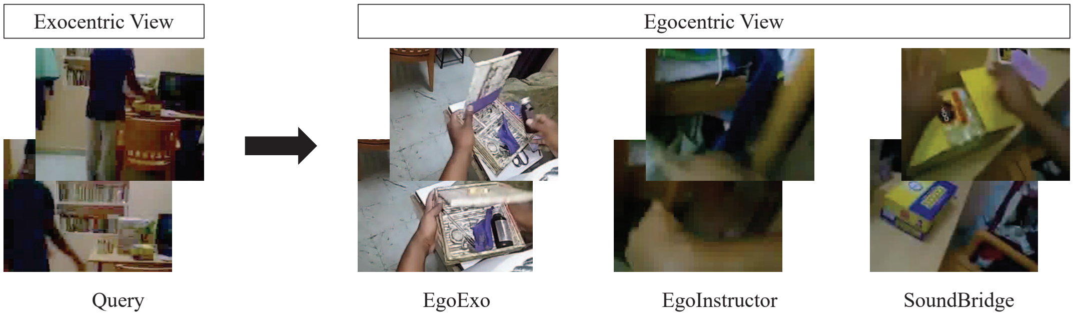


Figure 4. Visualization of an Exo-to-Ego video retrieval example, compared with EgoExo [12] and EgoInstructor [34], on an exocentric video query looking for *a person throws a box of groceries onto the table. The person opens the box and takes out an item.*

method differentiates between the actions “pick-up” and “put down” using sound cues. The pick-up action produces a soft rustling sound, whereas the put-down action generates a clunk sound when interacting with the iron bowl. Our proposed sound-aware modules effectively encode these audio features and their semantic descriptions (e.g., [pick up ginger] (soft rustling)). As a result, our method correctly ranks the ground-truth exocentric video in the top-1 position.

Figure 3 visualizes an Ego-to-Exo video retrieval example, comparing our method with EgoExo [12] and EgoInstructor [34]. The egocentric video depicts a person working on a vacuum, sneezing, and picking up a medicine bottle, while the visual cue for the sneeze is barely visible in the egocentric video. All methods successfully rank an exocentric video of a person with a vacuum in the top-1 position. However, EgoExo and EgoInstructor fail to capture either the sneezing action or the action of reaching for the medicine, leading to incorrect top-1 results. In contrast, our method effectively detects the sneeze and the action of picking up the pill bottle using sound cues, thereby successfully retrieving the ground-truth exocentric video. Similarly, Figure 4 visualizes an Exo-to-Ego retrieval example with a query for: *a person throws a box of groceries onto the table. The person opens the box and takes out an item.* All methods successfully retrieve videos that capture the actions of opening the box and taking out items. However, only our SoundBridge method correctly identifies the action of throwing the box onto the table, leveraging the thud sound as an essential cue.

### 5.3. Results on Text-Video Retrieval

The proposed sound-aware Ego-Exo representation space also enables text-video retrieval. Therefore, we compare its performance with baselines on both ego-(exo-)video-to-text and text-to-ego-(exo-)video retrieval tasks. Table 2 demonstrates that our method significantly enhances text-video retrieval performance on Ego-Exo datasets. Specifically, our method substantially improves the performance of the pre-trained egocentric model LaViLa [36], and the pre-trained exocentric model InternVideo [32]. Furthermore, our method outperforms recent Ego-Exo models, namely EgoExo [12] and EgoInstructor [34], by approximately 30.50% and 23.47% in overall performance across both datasets. The main improvement over recent Ego-Exo models is observed in queries with multiple nearest-neighbor (NN) answers. By incorporating sound aware-

Table 2. Performance comparison on text-video retrieval task.

| Method | EgoExoLearn | | | CharadesEgo | | |
|---|---|---|---|---|---|---|
| | T2V | V2T | Avg | T2V | V2T | Avg |
| FrozenInTime [3] | 6.03 | 4.05 | 5.04 | 9.40 | 5.62 | 7.51 |
| LaViLa [36] | 34.87 | 40.18 | 37.53 | 43.74 | 41.67 | 42.71 |
| EgoVLP [17] | 19.09 | 26.10 | 22.60 | 45.51 | 43.21 | 44.36 |
| InternVideo [32] | 18.59 | 26.55 | 22.57 | 47.34 | 46.63 | 46.99 |
| EgoExo [12] | 58.64 | 58.73 | 58.69 | 48.00 | 45.21 | 46.61 |
| EgoInstructor [34] | 66.37 | 66.68 | 66.53 | 46.93 | 45.98 | 46.46 |
| SoundBridge | **71.41** | **73.14** | **72.28** | **64.90** | **63.60** | **64.25** |

ness, our method effectively ranks the ground-truth videos higher among similar nearest-neighbor results. For example, given a text query searching for exocentric videos of a person washing dishes in the kitchen, EgoExo retrieves videos of a person chopping potatoes, while EgoInstructor retrieves videos of a person stir-frying. In contrast, our method distinguishes the washing action among these nearest neighbors by leveraging water-running sounds, allowing the ground-truth video to be ranked higher than videos of chopping and stir-frying.

## 5.4. Results on Downstream Tasks

We also evaluate the effectiveness of the learned egocentric visual feature (i.e., $f_{sv}^{ego}$) for scene recognition and the learned exocentric visual feature (i.e., $f_{sv}^{exo}$) for action recognition. We evaluate our model under two settings: (1) fine-tuning, where the training and test sets come from the same dataset, and (2) transfer learning, where the model is trained on one dataset and tested on another. The report performances of other models are all with the fine-tuning setting.

Table 3 compares the top-1 recognition accuracy with baseline methods. Specifically, for the egocentric scene recognition task, our method improves upon the egocentric model LaViLa [36] and the Ego-Exo model EgoInstructor [34] by 23.42% and 9.78%, respectively. Figure 5 illustrates an example of scene recognition. Given an egocentric video query where only a hand wiping a table with a towel is visible, it is difficult to infer that the scene is a kitchen. Other methods fail to predict the correct scene label. For example, FrozenInTime [3] misclassifies the video as a living room, while LaViLa [36] predicts it as a bathroom. In contrast, our method correctly identifies the scene by successfully aligning the query video with the corresponding exocentric video using sound cues from multiple actions (e.g., walking to the table and leaving the room). The egocentric visual feature (i.e., $f_{sv}^{ego}$) learns environmental and global visual cues from the exocentric video, which contains a range hood, an oven, and a refrigerator, allowing it to infer that the scene is a kitchen. For exocentric action recognition, our method achieves an average accuracy of 56.25%, which is lower than VideoMAE [30], as the model is trained on a large-scale action dataset. Despite this, our approach benefits from the sound-aware Exo-Ego representation space, as shown in Figure 6. Predicting a person's action from an exocentric view is challenging due to occlusions caused by the human body and the person being far from the camera. InternVideo [32] misclassifies the query video as the action "put", while EgoInstructor [34] assigns a high probability to the action "pick-up". In contrast, our method aligns the query video with its corresponding egocentric view using the sound of washing dishes (e.g., water splashing) and correctly classifies the action as "washing" with the highest probability.

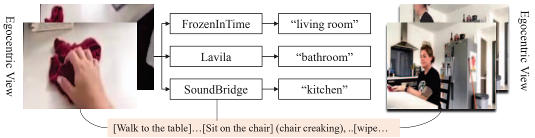


Figure 5. Visualization of the egocentric scene recognition results compared with FrozenInTime [3] and LaViLa [36].

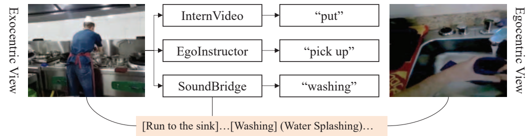


Figure 6. Visualization of the exocentric action recognition results compared with InternVideo [32] and EgoInstructor [34] .

Table 3. Result comparison on CharadesEgo scene recognition and EgoExoLearn action recognition. An asterisk (*) indicates the transfer learning setting, where a different dataset is used for training.

| Method | Ego-Scene | Exo-Action |
|---|---|---|
| FrozenInTime [3] | 8.25 | 6.45 |
| LaViLa [36] | 41.37 | 26.25 |
| EgoVLP [17] | 40.31 | 38.75 |
| InternVideo [32] | 43.00 | 45.00 |
| EgoExo [12] | 42.55 | 46.25 |
| EgoInstructor [34] | 46.51 | 47.50 |
| VideoMAEv2 [30] | - | **62.50** |
| InterImage [31] | 39.60 | - |
| SoundBridge* | 48.76 | 52.50 |
| SoundBridge | **51.06** | 56.25 |

## 5.5. Ablation study

### 5.5.1 Results on Different LLMs

We evaluated the impact of audio descriptions generated by different Large Language Models (LLMs) on the performance of cross-view video association tasks. Specifically, we used GPT-3.5 [4] and LLaMA3 [7] to generate audio descriptions using the same prompt. As shown in Table 5, the results indicate that while the cross-view video

Table 4. Ablation study on the proposed sound-vision and sound-text modules.

| Method | EgoExoLearn | | | | CharadesEgo | | | |
|---|---|---|---|---|---|---|---|---|
| | Eg2Ex | Ex2Eg | T2V | V2T | Eg2Ex | Ex2Eg | T2V | V2T |
| Baseline | 49.00 | 45.30 | 58.64 | 58.73 | 68.32 | 62.06 | 48.00 | 45.21 |
| Baseline + Sound-Text Module | 48.82 | 50.18 | 69.23 | 71.96 | 72.81 | 65.25 | 63.55 | 61.35 |
| Baseline + Sound-Vision Module | 48.46 | 47.73 | 69.10 | 73.09 | 74.35 | 66.55 | 63.59 | 61.11 |
| Baseline + Sound-Text&Vision Modules | **50.27** | **50.73** | **71.41** | **73.14** | **77.66** | **71.63** | **64.90** | **63.60** |

Table 5. Results with different LLMs.

| LLM | EgoExoLearn | | | CharadesEgo | | |
|---|---|---|---|---|---|---|
| | Eg2Ex | Ex2Eg | Avg | Eg2Ex | Ex2Eg | Avg |
| GPT-3.5 | 49.27 | 48.36 | 48.82 | 73.76 | 71.51 | 72.64 |
| LLaMA3 | **50.27** | **50.73** | **50.50** | **77.66** | **71.63** | **74.65** |

Table 6. Performance comparison of different anchors.

| Anchor | EgoExoLearn | | CharadesEgo | |
|---|---|---|---|---|
| | Eg2Ex | Ex2Eg | Eg2Ex | Ex2Eg |
| Origin Caption | 48.64 | 48.73 | 74.35 | 66.55 |
| Sound Caption | 47.09 | 49.91 | 56.97 | 51.42 |
| Sound Feature | 36.67 | 26.67 | 56.27 | 48.38 |
| Sound-Text Module | **50.27** | **50.73** | **77.66** | **71.63** |

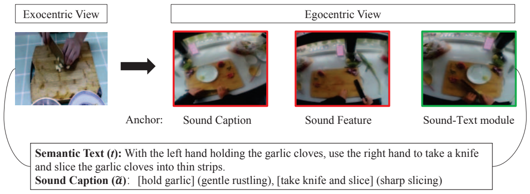
association performance with GPT-3.5 is lower than that achieved with LLaMA3, our method still surpasses existing approaches. This demonstrates the effectiveness of our proposed method.

### 5.5.2 Impact of the Sound-aware Modules

We conduct an ablation study on the proposed sound-vision and sound-text modules, using the state-of-the-art Ego-Exo model EgoExo [12] as the baseline. The performance of Ego-Exo video alignment and the overall accuracy of text-video retrieval are presented in Table 4. When using only the sound-vision or sound-text module, our method demonstrates competitive performance in the Ego-to-Exo video retrieval task. However, when both modules are jointly utilized, our method achieves significantly higher performance across the evaluated retrieval tasks.

### 5.5.3 Results of Different Anchors

We also investigate the effectiveness of using different features as anchors to align sound-aware visual representations. Specifically, we replace the $f_{st}$ in Eq. (11) and Eq. (12) with $f_t$, $f_c$ or $sum(f_a^{ego}, f_a^{exo})$. As shown in Table 6, using audio alone as an anchor is not ideal, especially for the short video dataset EgoExoLearn, where events heavily overlap and exhibit only slight variations. Upon analyzing the erroneous cases, we found that in many samples, the audio anchor alone fails to provide sufficient information to distinguish between videos from different views. As illustrated in Figure 7, a person is slicing garlic in the query video, and the text describes the semantic information of "a person slicing garlic". However, the sound features only convey the action "slice". During training, different-view videos were aligned solely based on the "slice" action, leading to incorrect retrieval of a video showing "slicing ginger" instead of "slicing garlic". Similarly, the sound caption lacks full contextual information and misinterprets the one-hand slicing action as two-hand slicing.



**Semantic Text ($t$):** With the left hand holding the garlic cloves, use the right hand to take a knife and slice the garlic cloves into thin strips.
**Sound Caption ($\tilde{a}$)**: [hold garlic] (gentle rustling), [take knife and slice] (sharp slicing)

Figure 7. An example of Exo-Ego video retrieval using different anchors in the SoundBridge model.

## 6. Conclusion

In this paper, we proposed two sound-aware modules as a bridge to address the large visual variance challenge in egocentric and exocentric video alignment. Both modules demonstrated significant improvements over state-of-the-art methods in Ego-Exo video retrieval and text-video retrieval tasks. Moreover, we verified that both egocentric and exocentric representations benefit from the sound-aware learning paradigm, achieving substantial improvements in downstream recognition tasks by leveraging complementary information from a correctly associated view. Furthermore, we found that our sound-vision and sound-text modules provide effective cues to distinguish the target video from its nearest neighbors based on sound.

## Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No.: 62372314, 62476097); the Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) (2023B1515120078); the Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (2024B1515040010).

## References

[1] Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the 1st Workshop and Challenge on Comprehensive Video Understanding in the Wild*, page 3, 2018. 2, 5

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 2

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 5, 7

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 7

[5] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, 2016. 2

[6] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022. 5

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 7

[8] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 6252–6261, 2019. 2

[9] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3293–3303, 2022. 2

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022. 1, 2

[11] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14867–14878, 2023. 3

[12] 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. 1, 2, 5, 6, 7, 8

[13] Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mohamed Omar. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12054–12064, 2023. 2, 3

[14] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10986–10994. IEEE, 2021. 2

[15] Patricia S Koskinen, Robert M Wilson, and Carl J Jensema. Using closed-captioned television in the teaching of reading to deaf students. *American Annals of the Deaf*, 131(1):43–46, 1986. 2, 3

[16] Haoxin Li, Yijun Cai, and Wei-Shi Zheng. Deep dual relation modeling for egocentric interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7932–7941, 2019. 2

[17] 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 Neu-*

*ral Information Processing Systems*, 35:7575–7586, 2022. 5, 7

[18] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *European Conference on Computer Vision*, pages 413–430. Springer, 2022. 2, 3

[19] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022. 2

[20] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Learning spatial features from audio-visual correspondence in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27058–27068, 2024. 3

[21] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 163–172, 2020. 2

[22] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19935–19947, 2022. 2

[23] 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. 5

[24] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978. 5

[25] Muhammad Bilal Shaikh, Douglas Chai, Syed Mohammed Shamsul Islam, and Naveed Akhtar. Multimodal fusion for audio-image and video action recognition. *Neural Computing and Applications*, 36(10): 5499–5513, 2024. 3

[26] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 2

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[28] 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*, pages 5250–5261, 2023. 2

[29] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13031–13040, 2023. 2

[30] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 5, 7

[31] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14408–14419, 2023. 5, 7

[32] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2024. 5, 6, 7

[33] Jianjia Xin, Lichun Wang, Kai Xu, Chao Yang, and Baocai Yin. Learning interaction regions and motion trajectories simultaneously from egocentric demonstration videos. *IEEE Robotics and Automation Letters*, 2023. 2

[34] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024. 1, 2, 5, 6, 7

[35] 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

[36] 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. 5, 6, 7