Ego-ImageBind: Correcting Alignment of Egocentric Videos

Evan Guenterberg
University of Maryland
eguenter@umd.edu

Xinchen Yu
University of Maryland
xyu054@umd.edu

Abstract

We present Ego-ImageBind, an optimized egocentric visual representation learning approach built upon ImageBind, aiming at improving egocentric visual downstream tasks particularly. By training a separate egocentric visual encoder, more detailed features of egocentric videos can be extracted, with the encoder performing better on noisy videos with rapid scene changing and partial captured limbs due to the nature of first-person perspective. Training is made possible with the publication of large-scale temporally aligned egocentric and exocentric videos from the Ego-Exo4D dataset. The code is available at https://github.com/derekthesnake/egocentric-encoding.git. (Some results can be added in abstraction when available) .

1. Introduction

Egocentric video understanding is an active topic in computer vision due to its importance in real-life scenarios, such as robotics and augmented reality. In these applications, collected data will necessarily be egocentric. While the egocentric nature of these visual data presents unique challenge for feature extraction, including rapid scene changing and partially captured limbs of the camera wearer. Since egocentric video differs from exocentric in camera motion and perspective of the subject, developing a dedicated model of understanding egocentric video is more effective than using an existing exocentric model [9, 12]. However, research has been productive in the exocentric multimodal realm: ImageBind [4] produced a set of encoders for 6 modalities with a shared latent space, and more recent work has been done integrating various modalities with large language models [3, 8]. It would be advantageous to use this established research with egocentric vision as well. In this paper we introduce an encoder for egocentric vision properly aligned to the ImageBind embedding space, that will allow us to do exactly that.

Until recently, this research was not possible. Existing egocentric vision datasets did not contain exocentric data points. Other works have attempted to transfer learnable features from exocentric video to egocentric video [7], but the lack of paired data made alignment very challenging. The release of Ego-Exo4D [6] introduced a large-scale dataset containing egocentric and exocentric recordings of the same events from multiple perspectives. See Fig. 1 for an example. Ego-Exo4D also contains many other modalities (point clouds, eye tracking, language descriptions, motion). This data allows us to train our egocentric-specific encoder to truly align egocentric and exocentric embeddings.

ImageBind [4] produced a set of encoders for 6 modalities with a shared latent space. 5 of the modalities (depth, audio, text, temperature, and motion) are aligned to the vision encoder. However, the vision data paired for each of these modalities is not entirely analogous. Audio, depth and thermal data was collected from datasets that are mainly exocentric, while IMU data was collected from Ego-4D [5], paired with egocentric video of the same events. This means that the IMU modality is aligned to a latent space different from the audio, depth, and thermal modalities.

Learning a suitable representation is crucial for downstream tasks, especially ones that rely heavily on multiple modalities. Previous works have proposed various tasks using egocentric video: action recognition, action anticipation, action segmentation, active object prediction, and more. Often these tasks have exocentric analogs with better performace, due to the abundance of exocentric datasets. Initial research is revealing that these tasks can benefit from extending to multiple modalities, indicating that this is the path forward in this field. [11] Thus to be able to leverage existing multimodal solutions for egocentric scenarios, egocentric video embeddings must properly align with other modalities, such as audio, that are independent of ego-/exocentrism.

2. Related Work

2.1. Egocentric Video Datasets

The number of large-scale egocentric video datasets has been increasing in recent years. EPIC-Kitchens[1, 2] are two large-scale egocentric datasets that record camera wear-

Figure 1. **Example frame from the Ego-Exo4D dataset.** Three exocentric perspectives are shown along with the egocentric perspective.

ers working in the kitchen. Ego4D[5] is a dataset published in 2021. It is a multi-modal dataset that includes audio, egocentric video, 3D pose, IMU, multi-channel audio, synchronized multi-person egocentric videos, and gaze data. However, not all modalities are present in every sample. Ego-Exo4D [6] is an upgraded version of Ego4D, including some other modalities, as well as time-synchronized exocentric video. To the best of our knowledge, no dataset published before the publication of Ego-Exo4D has large-scale temporally aligned egocentric and exocentric videos. The scarcity of data prohibited the alignment between egocentric and exocentric visual feature learning. Fig. 1 shows examples frame of the temporally aligned Ego-Exo4D dataset.

2.2. Contrastive Multi-modal Representation Learning

Contrastive learning achieves great success in extracting features from different modalities and projecting these features onto the same latent space. Pairs of related data points in different modalities should have embeddings that are similar to each other. Models like CLIP[10] utilize internet scale data to contrastively train encoders that embed image-text into the same latent space. ImageBind[4] further extends contrastive learning to 6 modalities by aligning each modality with visual. By unifying representations of different modalities in a common latent space, ImageBind works well as a backbone for different cross-modal downstream tasks.

2.3. Egocentric Video Understanding

Exocentric videos are well studied in various computer vision understanding tasks, including classification, object detection, object segmentation, and captioning. Egocentric videos, on the other hand, are mostly set in a "first person perspective", which differs from exocentric videos and presents unique challenges including rapid scene switching, primary actors located off-screen, and video continuity [12]. Directly adapting an existing Exocentric Video model would have a degraded performance, manifesting the importance of training a specialized model on egocentric video understanding.

Existing egocentric understanding techniques have attempted to address the lack of synchronized egocentric / exocentric videos by creating "pseudo-paired" data from separate egocentric and exocentric datasets [12]. This has had some success, but intuition suggests that a model would have better performance if the data it was trained on were properly synchronized.

[9] attempts to utilize the multi-modal nature of egocentric videos for captioning. In particular, egocentric video and IMU data are utilized. Our work does contrastive learning between egocentric videos and exocentric videos. By adapting the visual encoder of ImageBind[4], an emergent alignment would occur naturally between the egocentric video and text, audio, depth and thermal. IMU data can be trained with an egocentric encoder correctly aligned to the ImageBind latent space, generating a proper encoder for IMU.

Ego-Exo4D [6] presents more ego-exo video understanding tasks that require temporally and spatially aligned encoders, including the ego-exo correspondence and the ego-exo translation. The former task requires the model to identify and correspond the same object in both the egocentric and exocentric video, the latter task requires the model to predict the shape and position of the objects presented in one viewpoint but missed in the other. These tasks are rel-

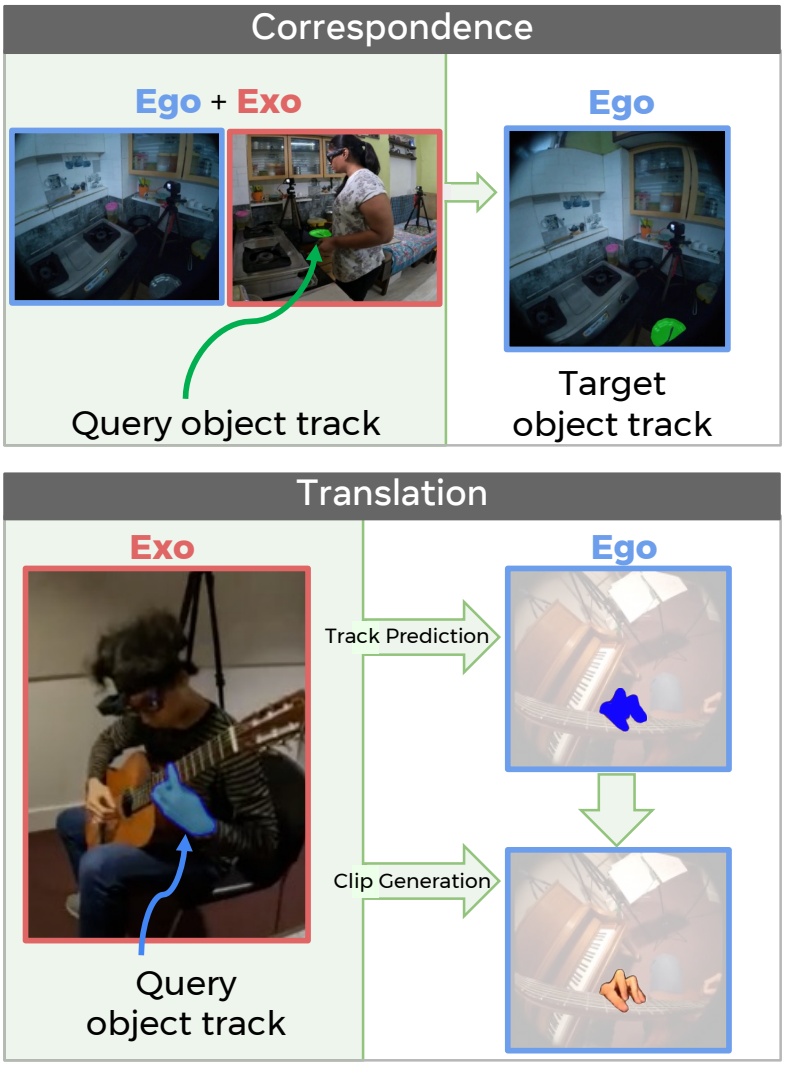

Figure 2. **Visualization of Ego-Exo Correspondence and Translation** Different viewpoint recording the same scene contains same objects. These objects may or may not be seen, while correspondence between these objects from videos of different viewpoints can be seen. Figure from Ego-Exo4D[6].

atively understudied because of the scarcity of temporally aligned egocentric and exocentric data. Our model provides a backbone for ego-exo understanding tasks through contrastive learning.

3. Approach

Our goal is to train a specialized video encoder that is capable of correctly understanding video under challenges presented by "first-person perspective" by aligning the embedding generated by egocentric videos to exocentric videos. We illustrate our approach in Fig. 3.

3.1. Egocentric Vision Encoder

We take a copy of the vision encoder as the new egocentric video encoder. Using the training split of the paired egocentric and exocentric videos, we perform contrastive learning to better align the egocentric video encoding to the ImageBind latent space.

We will follow the alignment procedure that ImageBind used: Given an exocentric video X_i and its paired egocentric video G_i, we encode them into normalized embeddings: $q_i = f(G_x)$ and $k_i = g(X_i)$ where f, g are visual encoders. The embeddings and the encoders are optimized using InfoNCE loss:

$$L_{X,G} = -\log \frac{\exp(q_i^\intercal k_i/\tau\ )}{\exp(q_i^\intercal k_i/\tau) + \sum_{j \neq i} \exp(q_i^\intercal k_j/\tau)}, \quad (1)$$

Video to video retrieval results on Ego-Exo4D							
	Ego to Exo			Exo to Ego			
	R@1	R@5	R@10	R@1	R@5	R@10	Ego/Exo cosine similarity
ImageBind	4.46%	17.66%	28.01%	5.44%	17.93%	28.27%	**0.738**
Ours (Entire Encoder)	0.09%	0.45%	0.98%	0.18%	0.45%	0.98%	0.037
Ours (Finetuning)	**8.56%**	**32.02%**	**48.17%**	**8.83%**	**32.02%**	**48.78%**	0.452

Table 1. **Retrieval task results.** Ego to Exo means that egocentric video was supplied, and exocentric video was retrieved. The result generated by updating the entire encoder is trained for one epoch, later epoch shows more severe overfitting. The finetuning result is trained for three epochs.

Audiovisual retrieval results on Ego-Exo4D							
	Video to Audio			Audio to Video			
	R@1	R@5	R@10	R@1	R@5	R@10	Audio/Ego cosine similarity
ImageBind Exo	1.19%	**7.39%**	11.81%	1.31%	**6.32%**	10.85%	**0.302**
ImageBind Ego	1.19%	5.13%	9.66%	0.95%	4.29%	8.11%	0.260
Ego-ImageBind (ours)	**1.43%**	6.92%	**13.12%**	**1.67%**	6.08%	**12.52%**	0.152

Table 2. **Cross-modal task retrieval task results.** Ego to audio means that egocentric video was supplied, and audio was retrieved. *ImageBind Exo* uses the standard ImageBind encoders and uses the exocentric video as a baseline. *ImageBind Ego* uses the standard ImageBind encoders and uses egocentric video to demonstrate the misalignment. Note that this cross-modal task was not trained for, and follows the ImageBind "emergent" alignment.

τ is a scalar temperature, and j is a non-matching pair (known as a 'negative'). ImageBind considers every example $j \neq i$ in the batch to be a negative, and we will too.

3.2. Data Processing

Egocentric and exocentric videos are preprocessed in the same way. Following ImageBind's approach, clips two seconds long are sampled uniformly from the video. From these clips, two frames are sampled from each. The frames are normalized and downscaled to 224x224.

Audio is extracted from the videos using ffmpeg. To match the video sampling, two second clips are sampled from the audio stream at corresponding locations. These clips are resampled to 16 kHz, transformed to Mel spectrograms, and normalized.

3.3. Implementation Details and Future Improvements

We directly adapt the Vision Transformer (ViT) used in ImageBind. We preserved the encoder for comparability and simplicity, while switching to other data representation or encoder may improve the egocentric video understanding ability, considering the quick scene change in video may need a denser key frame representation compared to exocentric videos. The misalignment between the IMU data and other modalities mentioned in Fig. 3 can also be corrected by retraining the IMU data encoder with the egocentric video encoder, after the alignment of the ego-exo video encoder. More modalities included in Ego-Exo4D [6] can also be added to the unified latent space, like gaze, point cloud, and pose. If we had more time we would do massive contrastive training for Ego-ImageBind to align these modalities as well.

However, due to the limited computation resource at our disposal, the encoder performance is negatively affected. With only 4 RTX A5000 GPUs, the batch size we can fit in the model is only 4 when updating the entire encoder, two orders of magnitude smaller than the encoders trained in ImageBind [4]. Updating only the projection head of the model allows the batch size to be increased to 128, which is still only one-fourth of the batch sizes used in ImageBind. The smaller batch size means that the InfoNCE loss function captures interactions between fewer data points, leading to inefficient training and poor performance.

Ego-Exo4d [6] can be utilized in a more efficient way than directly adapting the data preprocessing approach in ImageBind [4]. By trimming long Ego-Exo4d videos into shorter segmentation, more datapoints can be acquired. We also did not utilize the entire Ego-Exo4d [6] dataset. The dataset provides 4 exocentric cameras for each scene, while we only utilize one of such scene. Additionally, our training processed the original 4K videos for each training epoch, but immediately downsampled them before training. In the future, downsampling the videos beforehand and saving them to disk would greatly speed up training time by reducing the data processing overheads.

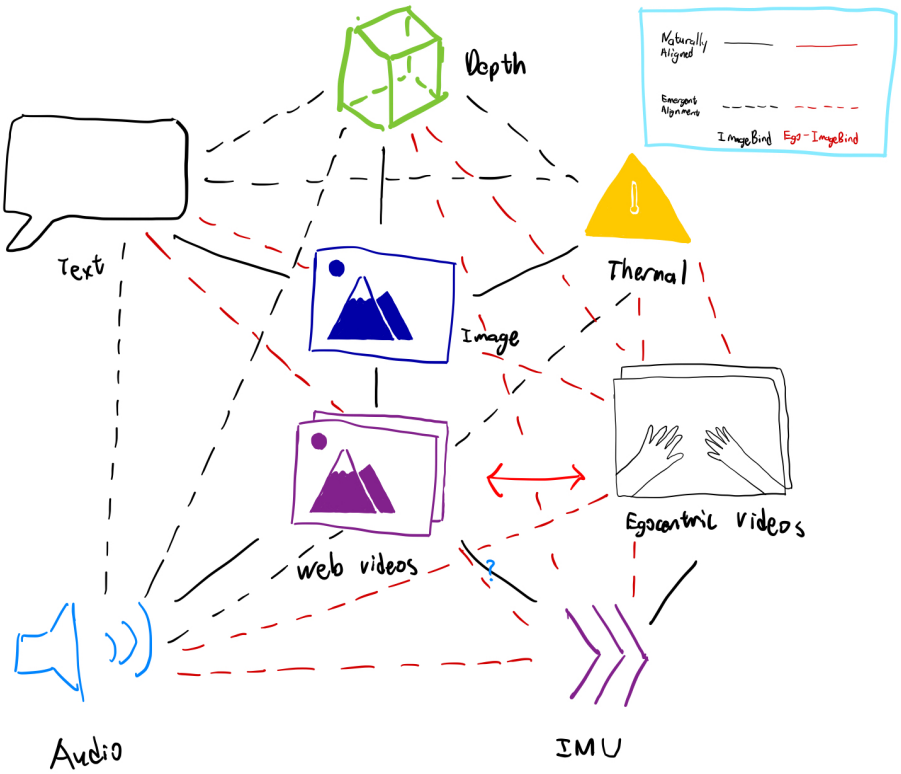

Figure 3. **Structure of Ego-ImageBind** ImageBind[4] correlates visual with 5 modalities: text, audio, depth, thermal, and IMU data. However IMU-Video alignment is trained with Ego4D[5] dataset. The different characteristics between egocentric videos and abundantly existing exocentric videos may cause the alignment between IMU data to misaligned with other modalities in ImageBind. The main goal of our work is to align egocentric video and exocentric video, thus allowing emergent alignment between egocentric video with other modalities. IMU data can also be correctly aligned as a result.

4. Experiments

4.1. Experiment Setup

The training is performed on the UMIACS server using 4 RTX A5000, data parallelism is used for training and inference. Following ImageBind [4], each video in Ego-Exo4D [6] is represented by 5 clips with 2 frames per clip. The generated embeddings per clip is averaged into one single embedding. Due to the limited computation resource, two hyperparameter settings are implemented.

The first setting updates the weight of the entire encoder. However, the maximum batch size that fits onto the GPU is 4. In addition, due to the strategy of representing each video with one single embedding, the number of data points we utilize is about 3000. Overfitting is severe in this setting. The results can be seen in Section 4.2

The second setting updates only the encoder projection head. With fewer gradients needed to be traced, a batch size of 128 is used. Fewer weight updates per epoch slows down overfitting. With less weight being changed, the learning capacity of the model is also reduced, thus a more general representation may be learned.

4.2. Retrieval Task Results

To quantify the effectiveness of the alignment, we perform various retrieval tasks with the test split. To start, we retrieve the egocentric video given the exocentric video and vice versa using the original ImageBind vision encoder for both. We perform the same task, but using our fine-tuned egocentric vision encoder for the egocentric videos. We also capture the average cosine similarity score for ego/exo pairs as a method of measuring the alignment of the latent spaces. The results are shown in Table 1.

We also perform retrieval across modalities: audio from egocentric video and vice versa. Again, this is compared to the ImageBind encoder as a baseline to validate our egocentric specialized encoder. This gives a much better idea of whether the alignment we induce is useful. The results for this are shown in Table 2. Updating the entire encoder proved to be an ineffective method of fine-tuning the vision encoder, inducing highly negative ego/exo retrieval results. For this reason, we do not attempt cross-modal re-

trieval with this encoder. The finetuned encoder performs better than audio retrieval using exocentric video with exocentric encoder, which may suggest the comparison is still unfair. Finetuned exocentric video may learn better feature retrievals using the extra information provided by Ego-Exo4D[6] data. More experiments on other datasets may be necessary in the future.

5. Discussion and Limitation

The fine-tuned egocentric encoder vastly outperformed the original ImageBind encoder in unimodal retrieval, indicating successful alignment of the egocentric vision space to the exocentric vision space. R@1 and R@5 accuracy nearly doubled, and R@10 accuracy jumped approximately 70%. Surprisingly, the cosine similarity score decreased from the original ImageBind encoder. The mechanism for this is not known and requires further investigation.

For cross-modal retrieval, it is clear that the egocentric video when used with exocentric encoders produces a drop in accuracy. However, our fine-tuning was able to restore the accuracy. It is heartening to see that in R@5, the exocentric video retrieval even beats our accuracy, indicating that our training probably does not overfit the egocentric encoder to this dataset, but this cannot be confirmed without further study. Again we can see that the cosine similarity score declined in our fine-tuning, but we do not understand this result.

Our results show the misalignment between the embedding of egocentric and exocentric videos generated by ImageBind [4] vision encoder, suggesting that different encoders may be needed for encoding egocentric and exocentric videos. This indicates that the IMU encoder for ImageBind should be retrained, since the existing encoder aligns IMU data with the ImageBind vision encodings of egocentric data. By aligning instead to our corrected egocentric vision encoder, the IMU encoder would be more accurate for cross-modal tasks.

We provide an egocentric encoder trained with data from Ego-Ego4D [6] that shows the potential of generating embeddings of egocentric videos which are aligned correctly with ImageBind [4] latent space.

6. Contributions

Evan Guenterberg wrote the code to load the data and to perform the retrieval tasks. Xinchen Yu wrote the training code and model architecture. Both contributed to the writing of the paper.

References

[1] Dima Aldamen, Davide Moltisanti, Evangelos Kazakos, Hazel Doughty, Jonathan Munro, William Price, Michael Wray, Tobias Perrett, and Jian Ma. Rescaling egocentric vision. In *IJCV*, 2021. 1

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[3] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. 1

[4] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 1, 2, 4, 5, 6

[5] 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, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, 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 *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 5

[6] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun So-

mayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shout, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, 2024. 1, 2, 3, 4, 5, 6

[7] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos, 2021. 1

[8] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024. 1

[9] Katsuyuki Nakamura, Hiroki Ohashi, and Mitsuhiro Okada. Sensor-augmented egocentric-video captioning with dynamic modal attention. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 4220–4229, New York, NY, USA, 2021. Association for Computing Machinery. 1, 2

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Icml*, 2021. 2

[11] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–20, 2022. 1

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