Learning Temporal Video-Language Grounding for Egocentric Videos

Priyanka Mandikal **Alex Chandler** **Liyan Tang**
Department of Computer Science, The University of Texas at Austin

Abstract

In this paper, we address the challenge of learning temporal structure in long-horizon egocentric videos by pretraining a large video-language model on the Ego4D dataset. Our goal is to transfer the pretrained model to downstream video understanding tasks such as video-text retrieval. To achieve this, we propose a novel contrastive learning approach that combines hard negative mining via nearest neighbor sampling with temporal clip reordering. Our key insight is to use a sampling strategy that incorporates strong and weak samples in the contrastive learning objective to encourage learning temporally robust video-language representations. We also develop a temporal reordering objective to unshuffle and reorder video clips and their corresponding narrations. Our approach outperforms existing state-of-the-art methods on the challenging EgoMCQ benchmark, demonstrating a significant improvement in both intra-video and inter-video metrics.

1 Introduction

Spatio-temporal comprehension of egocentric video data is vital for addressing various downstream computer vision tasks, with applications ranging from augmented reality to providing robots with expert demonstration data. Consider a person wearing AR glasses while performing household chores. A video understanding model may need to answer questions like "Where did I leave my keys?" or "Did I turn off the stove before leaving the house?". Responding to such queries necessitates a deep understanding of both natural language queries and a technique for mapping them to specific video segments. We are motivated by this intricate yet highly valuable task of answering natural language queries derived from egocentric videos.

In this work, we investigate the challenge of effectively grounding video-language embeddings to comprehend the temporal structure of long-horizon egocentric videos. Egocentric videos, captured using wearable cameras, provide a unique first-person perspective of the surrounding environment. Our objective is to pretrain a large-scale video-language model on extensive egocentric video datasets, enabling its transfer to a variety of downstream video understanding tasks (Fig. 1).

To solve this problem, we consider the large-scale Ego4D dataset(Grauman et al., 2022), which encompasses 3000 hours of egocentric videos featuring humans engaged in various day-to-day activities. These videos are accompanied by rich, fine-grained natural language narrations. We utilize the EgoClip subset from EgoVLP (Lin et al., 2022b), comprising a pretraining set of 3.8M video-text pairs derived from Ego4D. Our aim is to pretrain a robust multi-modal video-language model on EgoClip, enabling efficient performance on the downstream task of answering natural language queries. We evaluate our approach on the EgoMCQ benchmark from EgoVLP (Lin et al., 2022b), modeled as a video-text retrieval task. Given a natural language narration and five video segments, the task involves matching the narration to the appropriate segment. We strive to develop effective contrastive learning objectives for training large video-language models to excel at video-text retrieval tasks involving extensive egocentric videos.

Pretraining large-scale video-language models for egocentric understanding involves several challenges: **(1)** Egocentric videos are inherently complex, exhibiting significant diversity in actions, environments, and camera movements. **(2)** Long-horizon videos demand reasoning over expansive contexts, making it difficult to pinpoint relevant information within multi-segment videos. **(3)** Natural language queries can be ambiguous, necessitating a multi-modal understanding of both visual and linguistic information. Mapping queries to specific video segments requires fine-grained spatio-temporal reasoning and an understanding of context and

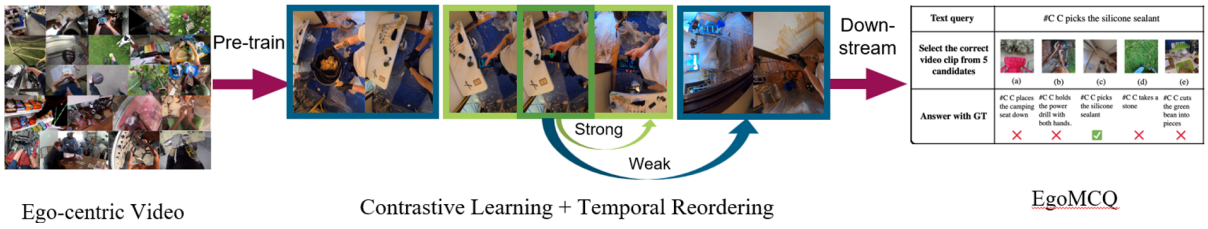


Figure 1: **Egocentric video-language pretraining for downstream tasks.** We perform large-scale video-language pretraining on egocentric videos from EgoCLIP for the downstream task of video-text retrieval on EgoMCQ.

temporal relationships.

To tackle these challenges, we propose a novel contrastive learning approach that incorporates hard negative mining and temporal reordering of egocentric video clips to enhance performance on the downstream EgoMCQ video-text retrieval task. We integrate both strong and weak contrastive learning with temporal reordering during pretraining. Our key insight involves a novel sampling strategy that combines strong and weak samples in the contrastive learning objective, fostering the development of temporally robust video-language representations. Unlike prior works such as EgoVLP (Lin et al., 2022b), which employ random sampling during contrastive learning, our method utilizes a 50-50 strong-weak sampling strategy to mine hard negatives. Additionally, we introduce a temporal reordering objective for unshuffling and reordering clips and their corresponding narrations.

Through experiments on 10% of the EgoCLIP dataset (380k video-text pairs), our approach outperforms EgoVLP on the challenging EgoMCQ benchmark, achieving a significant improvement in both intra-video and inter-video metrics. This highlights the effectiveness of our proposed contrastive learning approach, which combines hard negative mining and temporal reordering for grounding video-language embeddings in large-scale egocentric video datasets.

The primary contributions of this work are as follows:

- We propose a novel contrastive learning approach that combines hard negative mining with temporal reordering of egocentric video clips, promoting improved performance on the EgoMCQ video-text retrieval task.

- We introduce a new sampling strategy that integrates strong and weak samples in the contrastive learning objective, fostering the development of temporally robust video-language representations for large-scale egocentric video understanding.

- We demonstrate the effectiveness of our approach through experiments on a subset of the EgoCLIP dataset, outperforming existing state-of-the-art methods on the EgoMCQ benchmark and achieving significant gains in both intra-video and inter-video metrics.

Our findings provide valuable insights into the challenges and opportunities of pretraining large-scale video-language models for egocentric understanding, paving the way for further research and development in this domain.

2 Related Work

2.1 Video Language Pretraining

Video-language pretraining aims at improving the representation of videos and their associated textual information such as captions and transcriptions. The learned representations can be leveraged to effectively perform downstream tasks such as video-text retrieval (Zhu et al., 2023), video question answering (Zhong et al., 2022), and video captioning (Wang et al., 2018). Popular video-language pretraining tasks include Masked Language Modeling (MLM), Video Language Matching (VLM), and Frame Order Modeling (FOM). MLM, inspired by BERT's pretraining objective (Devlin et al., 2019) has been adapted for video-language pretraining by masking textual information (Zhu and Yang, 2020), video features (Peng et al., 2022; Bao et al., 2022; Xie et al., 2022) or both simultaneously (Li et al., 2022; Sun et al., 2019). VLM focuses on learning the correspondence between a video and its related text, thereby improving alignment and understanding (Zellers et al., 2021; Zhang et al., 2021; Li et al., 2021). FOM helps develop the skill of temporal reasoning by predicting the original order of a sequence of shuffled frames within either the visual or both visual and textual modalities (Zellers et al., 2021; Lei et al., 2021; Li et al., 2020). In

this work, we explore how we can best combine both FOM and VLM techniques by combining both contrastive learning on weak and strong pairs with temporal reordering.

2.2 Video-Text Retrieval

Video-text retrieval (VTR) is the task of locating and identifying specific video segments or moments that match the given natural language query or description. Typically, VTR systems involve two main steps: extracting spatial and temporal visual features and textual information from the input video and query, respectively; and establishing a correspondence between them to identify relevant video segments. Recent work relies on Transformer-based methods to extract spatial and temporal features from videos (Dosovitskiy et al., 2021; Radford et al., 2021; Bain et al., 2021; Bertasius et al., 2021) and textual information from sentences (Sanh et al., 2019; Devlin et al., 2019). Establishing correspondence between video and textual feature representation can be further divided into global and local matching. Global matching focuses on capturing the overall semantic similarity between the video and textual representation (Zhu and Yang, 2020; Yang et al., 2020). Local matching, instead, focuses on the similarity between fine-grained local features extracted from the video frames and textual description (Zhao et al., 2022; Cheng et al., 2021). Our work falls into the global matching category, where we use contrastive learning to learn stronger global representations of visual and textual features.

2.3 Egocentric Video Datasets

Egocentric datasets, comprising videos and images captured from a first-person perspective, are crucial for developing applications in the fields of robotics, human-computer interaction, and augmented reality. Several datasets have been introduced to address the unique challenges posed by the egocentric perspective, such as object occlusions, motion blur, occlusions, and varying lighting conditions. Ego4D (Grauman et al., 2022) is a comprehensive dataset encompassing 3,670 hours of videos with dense textual narrations covering a diverse array of daily-life activities. The EPIC-Kitchens dataset (Damen et al., 2018) is another popular dataset, featuring 100 hours of egocentric recordings in 45 kitchens, accompanied by multi-language narrations. Other notable egocentric datasets include the GTEA (Li et al., 2015), ADL (Pirsiavash and Ramanan, 2012), and UTE (Sinha et al., 2015) datasets, which have also been utilized for various egocentric video analysis tasks. These datasets facilitate the development of models that understand the visual world from the viewpoint of the person capturing the data. We selected Ego4D for our study due to its large-scale and diverse data, ensuring that the learned embeddings are generalizable across a wide range of real-world scenarios. Furthermore, the multi-modal annotations and benchmarks provided by the Ego4D dataset enable effective comparison of our pretraining objectives with existing state-of-the-art methods.

3 Approach

Our approach aims at training a model to learn a good feature representation of video-language by employing a two-fold strategy, as illustrated in Fig. 2: (1) Contrastive learning: provide the model with strong and weak negatives of neighboring video clips from the same video, and (2) Temporal Reordering: unshuffle the neighboring video clips from the same video. We describe each of these techniques in the subsequent sections.

3.1 Hard Negative Mining for Video-Language Contrastive Learning

Contrastive learning is a widely used approach in deep learning that seeks to learn an encoding function capable of mapping positive samples closer to the anchor in the embedding space while simultaneously driving negative samples farther away from the anchor. To achieve this, we have adapted the EgoNCE loss function proposed in Lin et al. (2022b).

$$\mathcal{L}_{\text{v2t}}^{\text{ego}} = \frac{1}{|\widetilde{\mathcal{B}}|} \sum_{i \in \widetilde{\mathcal{B}}} \log \frac{\sum_{k \in \mathcal{P}_i} \exp\left(\mathbf{v}_i^T \mathbf{t}_k / \tau\right)}{\sum_{j \in \mathcal{B}} \left(\exp\left(\mathbf{v}_i^T \mathbf{t}_j / \tau\right) + \exp\left(\mathbf{v}_i^T \mathbf{t}_{j'} / \tau\right)\right)}$$

EgoNCE In EgoNCE, positive samples set $\mathcal{P}$ are constructed by grouping samples within a batch $\mathcal{B}$ that share similar verbs or nouns in their respective narrations. This enables the model to learn to identify semantic similarities between samples. On the other hand, negative pairs are constructed using a hard negative mining strategy, where samples that have different actions within the same video are selected as negative samples. The objective can

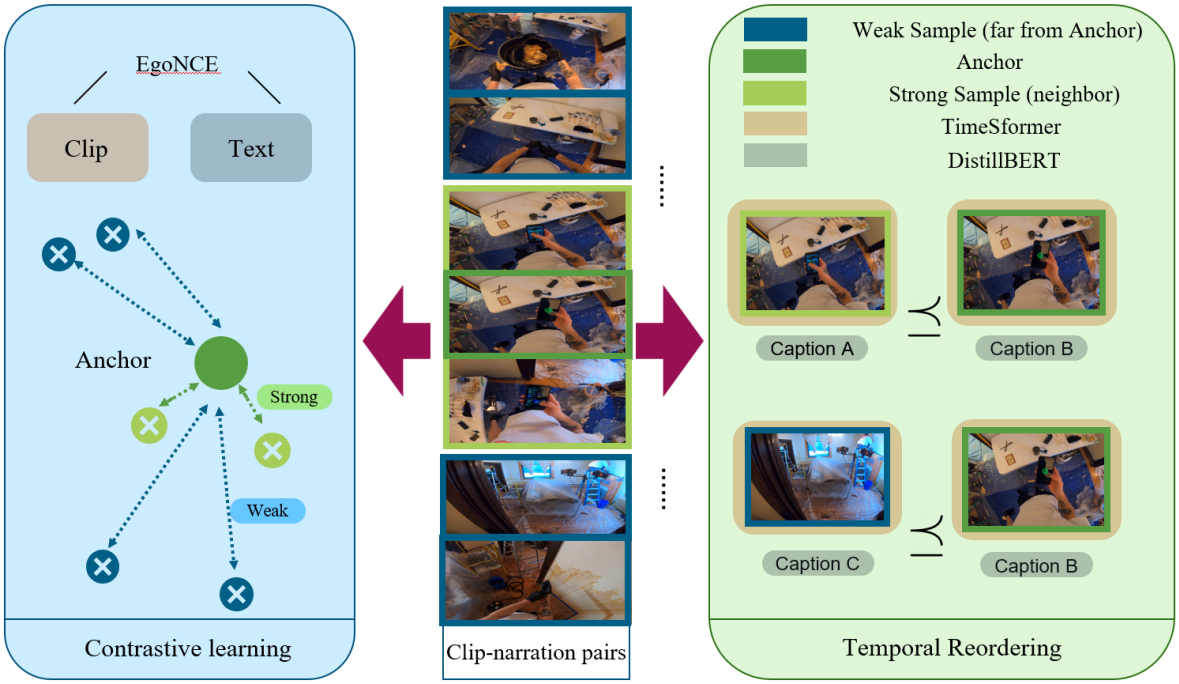


Figure 2: **Overview of our approach.** We incorporate strong and weak hard negative mining via nearest neighbor sampling along with temporal clip reordering to perform video-language pretraining.

be formulated as Equation (3.1), where $\widetilde{\mathcal{B}}$ consists of all samples in $\mathcal{B}$ and a hard negative sample for each sample, respectively (i.e., $|\widetilde{\mathcal{B}}| = 2|\mathcal{B}|$). $\mathbf{v_i}$ and $\mathbf{t_j}$ represents the embedding for video v and text t. However, a potential issue with this approach is that negative samples may be too dissimilar from positive samples, rendering the contrastive learning process ineffective.

Adaptation To address this issue, we propose a novel negative sampling strategy that utilizes nearby video segments from the same video as negative samples to construct $|\widetilde{\mathcal{B}'}|$. By selecting neighboring clips located one or two clips away from the anchor, we ensure that the negative samples in $|\widetilde{\mathcal{B}'}|$ have different actions while still being close enough to the anchor to make the learning process effective. This approach forces the model to differentiate between subtle differences in the visual field, leading to better overall performance.

3.2 Temporal Clip Reordering

Temporal reordering is motivated by the fact that videos are inherently temporal in nature, and learning to reason about temporal relationships between video clips is crucial for many video-related tasks. Although prior work has not explicitly explored the temporal unshuffling of video segments, our work is inspired by methods such as MERLOT (Zellers et al., 2021), which seek to extend the concept of temporal loss from individual frames to sequences of video clips. Building on this concept, our second objective is to introduce a novel temporal reordering loss that encourages the model to learn temporal reasoning by unshuffling sequences of video clips.

Previous research in this area has typically involved randomly selecting video clips and attempting to reorder the shuffled sequence. However, these approaches often fail to consider the relationships between the selected video clips. For instance, if the sampled video clips are far apart from each other and lack an obvious connection, reordering these clips may not be meaningful, as even humans would struggle to correctly order them without additional information and context.

To address this limitation, we take into account the contextual relationships between video clips. Specifically, we leverage contextual information to identify which video clips are likely to be related by ordering nearby video clips, wherein the objective now is to compute similarity scores between video-language embeddings of neighbors and reorder clips such that m-step neighbors are closer together in the embedding space compared to n-step neighbors, where $m > n$. This enables our model to reason about the temporal relationships between video clips in a more meaningful and contextually relevant way. Additionally, it allows us

to intuitively restructure the embedding manifold with a richer spatio-temporal prior, thereby improving the performance of our downstream task of video-text retrieval.

4 Experiments

4.1 Implementation Details

In our experiments, we train on 10% of the EgoCLIP dataset, conducting five epochs per experiment. We use a batch size of 8 and a learning rate of 3×10^{-5} to optimize memory usage and computational efficiency. To increase the diversity and robustness of our model, we apply various image augmentations at the video level, including random cropping, color jitter, rotations, and other transformations. We preprocess video frames by resizing them to a resolution of 224 $\times$ 224 pixels and employ patches of size 16 in the vision transformer for capturing fine-grained details. Our model utilizes the TimeSFormer architecture as the vision backbone, designed for efficient spatiotemporal data processing. We also incorporate DistillBERT for text embedding, ensuring effective multimodal learning through the encoding of textual information. We implement distributed training across eight NVIDIA RTX 6000 GPUs to manage computational resources and memory requirements effectively. This setup allows us to complete the training process for each experiment within a 24-hour timeframe, facilitating rapid iteration and advancement of our research.

4.2 Dataset

We train our model on an evaluation set from the EgoCLIP dataset, a subset of the Ego4D dataset used for pretraining video-language models, consisting of 3.8M video-text pairs, provided by (Lin et al., 2022a). EgoCLIP is a 1st-person video-text covering a diverse domain of daily human activities. EgoCLIP benefits from dense timestamp-level narrations from two separate narrators. EgoCLIP averages 21.9 clips per minute with an average clip duration of 1.0 seconds and a standard deviation of 0.9 seconds. We visualize sample clips with their associated narrations in Fig. 3.

4.3 Evaluation Metrics

We evaluate the proposed method on the Egocentric Multiple-Choices-Question (EgoMCQ) benchmark provided by EgoVLP (Lin et al., 2022b). EgoMCQ is designed for evaluating video-text alignment in egocentric videos. It contains questions created from the Ego4D dataset and focuses on assessing the ability of models to understand and align the spatio-temporal information present in the videos with corresponding textual descriptions. We visualize samples from the EgoMCQ validation set in Fig. 4.

There are two evaluation settings for the dataset. In the **inter-video** setting, each multiple-choice question consists of five video clips taken from different videos. This setting aims to test a model's ability to distinguish instances from different scenarios. In contrast, the **intra-video** setting presents a more challenging task by grouping five continuous clips together. This setting focuses on fine-grained context clues and serves as a specific form of video-text localization. The inter-video and intra-video settings consist of 39k questions covering 198k narrations and 468 hours of video.

We evaluate the model performance via accuracy, i.e., whether the model can select the correct answer from multiple choices. The intra-video setting is particularly challenging as the intra-video accuracy is around 50% from Lin et al. (2022b). We aim to design better training methods that allow models to better understand fine-grained context clues in egocentric videos.

4.4 Comparison

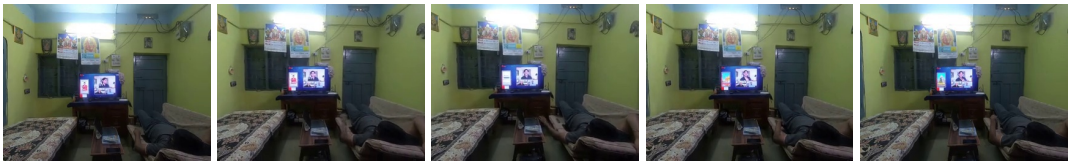
We first create a baseline that tests the model given by the EgoVLP paper with their contrastive loss function, trained on only 10% of the data, where negative samples are sampled randomly from the same video. We then run our unique sampling strategy that uses both strong and weak sampling. Finally, we experiment with our own temporal reordering function, using only our strong sampling technique. This approach is motivated by the goal of selecting video clips that are proximal to one another and thus more likely to be temporally related, forcing the model to learn their temporal ordering. Conversely, if we selected far-away clips, it would be unrealistic to expect the model to order clips without giving the model more intermediate clips or an additional input modality.

Our proposed model extends and improves upon EgoVLP by focusing on a better understanding of the temporal structure and fine-grained context clues in egocentric videos. By introducing a harder negative sampling objective along with a temporal reordering objective, we expect our model to out-

(a) `ties the vegetable with a band.`

(b) `moves the right hand.`

(c) `C stretches his left hand.`



(d) `A man moves hand from the table.`

Figure 3: **Visualization of clip-text pairs.** We visualize five frames sampled uniformly for each clip and take its narration as its caption.

perform EgoVLP, particularly in more challenging tasks such as the intra-video setting of the Egocentric Multiple-Choice Question (EgoMCQ) benchmark.

4.5 Results

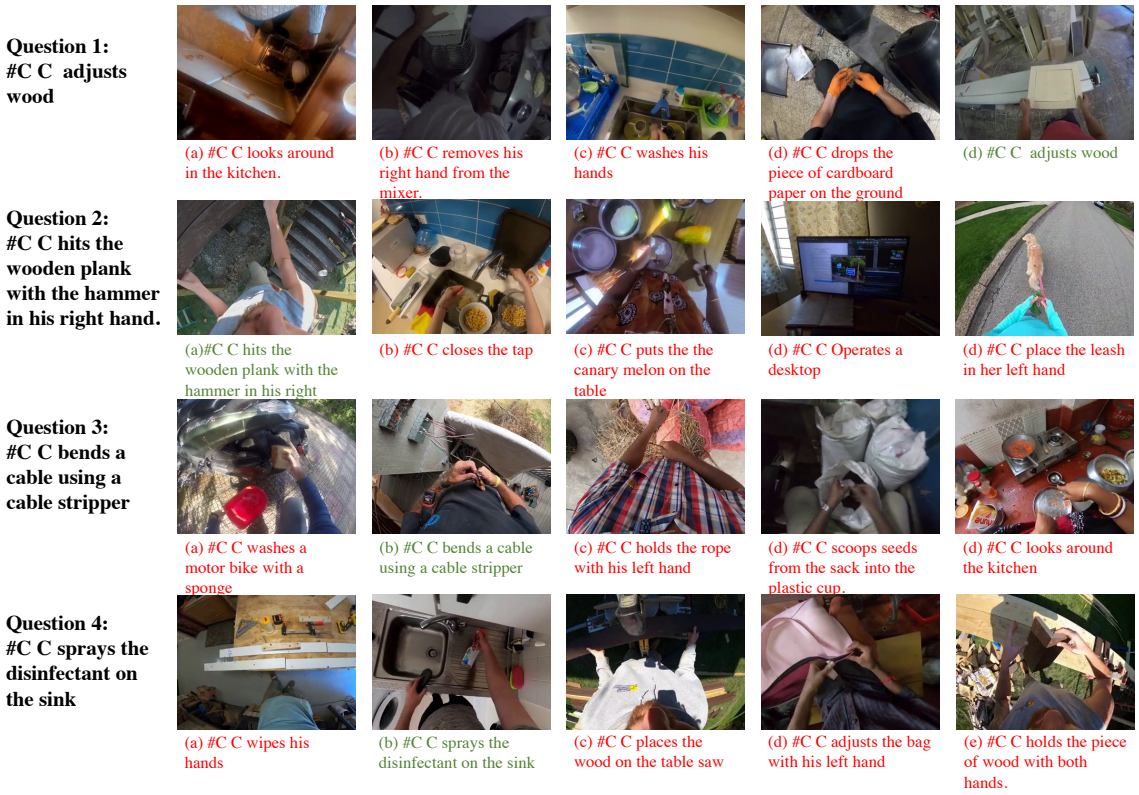
In Table 1, we present the experimental results comparing our method with the EgoVLP baseline for different training settings. The table compares the performance of the methods using only 10% of the training data, under varying epochs, sampling strategies, and temporal reordering conditions. The EgoMCQ accuracy is reported for both intra-video and inter-video scenarios.

Focusing on the results for the models trained on 10% of the data, we observe that our method outperforms the EgoVLP baseline in both intra-video and inter-video accuracy. Specifically, when comparing our method with the EgoVLP baseline trained for 5 epochs using the random sampling strategy, our strong plus weak contrastive learning method achieves an inter-video accuracy of 84.0% and an intra-video accuracy of 47.3%, while the EgoVLP baseline reports an inter-video accuracy of 82.9% and an intra-video accuracy of 46.2%.

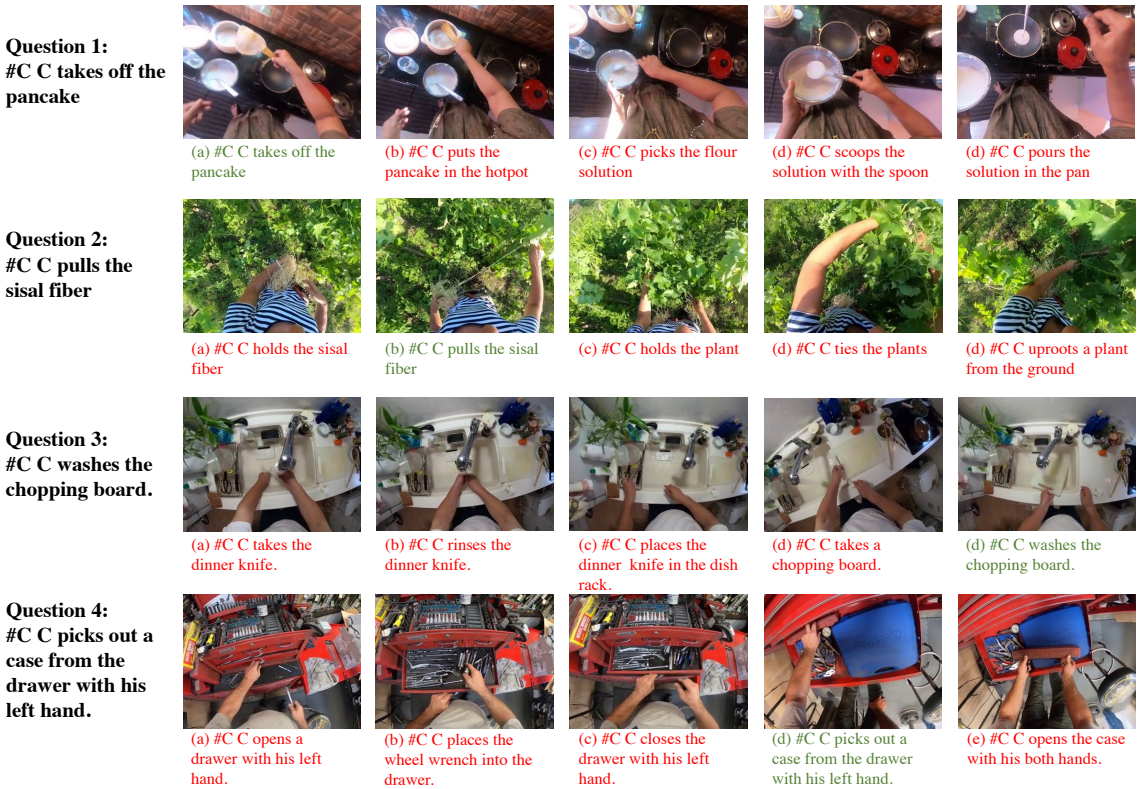
Furthermore, our method, when employing a strong sampling strategy with temporal reordering, achieves an inter-video accuracy of 84.3% and maintains the same intra-video accuracy (47.3%) as our method without temporal reordering. This suggests that our proposed method is more robust and effective in handling the challenging NLQ task on egocentric videos, even when using a reduced amount of training data.

5 Discussion

In this study, we proposed a novel approach to learning the temporal structure of long-horizon egocentric videos by pretraining a large video-language model on the Ego4D dataset. Our method demonstrated promising results, outperforming existing SOTA methods on the challenging EgoMCQ



(a) Inter-video setting



(b) Intra-video setting

Figure 4: Visualization of EgoMCQ under two settings. Left are the text questions; Right are the five candidate clips for each question and the text below as clip's narrations. The correct clip's narration is highlighted in green and the wrong in red.

Method	Training Data	Epochs	Sampling	EgoMCQ Acc (%)	
				Inter-Video	Intra-Video
InfoNCE	100 %	10	Random	89.4	51.5
EgoVLP (NeurIPS '22)	100 %	10	Random	90.6	57.2
EgoVLP (NeurIPS '22)	10 %	5	Random	82.9	46.2
Ours	10 %	5	Strong + Weak	84.0	**47.3**
Ours w/ reordering	10 %	5	Strong	**84.3**	**47.3**

Table 1: **Performance on the EgoMCQ benchmark provided by EgoVLP (Lin et al., 2022b)** Our strong + weak sampling strategy for contrastive learning outperforms EgoVLP in both intra-video and inter-video accuracy when trained on 10% of the EgoCLIP training data. Additionally, adding temporal reordering to the loss objective further improves accuracy in the inter-video setting.

benchmark. However, there are some limitations to our approach, and we identify possible future directions to improve upon the current work.

5.1 Strengths

Our two-fold approach to training combines both hard negatives of neighboring video clips and a temporal reordering objective within the contrastive loss. This strategy significantly improves the model's ability to learn the temporal ordering of video-language embeddings, leading to better performance on downstream tasks. Moreover, the adoption of the TimeSFormer architecture enables a fair comparison to benchmark performances on downstream tasks in the EgoVLP framework.

5.2 Limitations

Despite our model's ability to reorder temporal sequences, it may still be unable to capture the long-term temporal dependencies required for a comprehensive understanding of the egocentric activities in the videos. Additionally, we were unable to train our model on the full EgoVLP dataset due to the computational demands posed by the TimeSFormer architecture.

5.3 Future Work

Several potential avenues for future research can be considered to build upon our current work:

1. Further downstream evaluation: Evaluate the performance of our pretrained model on additional downstream tasks, such as natural language queries (NLQ), visual question answering (VQA), and moment query tasks, to establish its generalizability and utility.

2. Improved temporal reordering: Investigate more sophisticated temporal reordering techniques to enhance the model's ability to capture complex temporal dependencies between actions in egocentric videos.

3. Hyperparameter tuning: Conduct an extensive search for optimal hyperparameters to maximize the performance of the model on the EgoMCQ benchmark and other downstream tasks.

4. Explore newer video transformers: Investigate the potential benefits of incorporating more recent video transformer architectures, such as VideoMAE and VideoMAEv2, to further improve the model's ability to learn temporal structures in long-horizon egocentric videos.

By addressing these limitations and exploring future directions, we aim to further advance the field of egocentric video understanding and its applications in robotics, augmented reality, and human-computer interaction.

6 Conclusion

In this paper, we presented a novel approach to learning the temporal structure of long-horizon egocentric videos by pretraining a large video-language model on the Ego4D dataset. Our two-fold method, combining hard negative mining and temporal clip reordering, effectively improves the model's ability to learn temporally robust video-language embeddings. Our approach demonstrates significant performance gains on the EgoMCQ benchmark, outperforming state-of-the-art methods, such as EgoVLP. Future research will focus on refining our method and exploring newer architectures to further improve performance on complex, real-world scenarios from an egocentric perspective.

References

Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*.

Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. 2021. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*.

Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. 2021. Understanding chinese video and language via contrastive multimodal pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online. Association for Computational Linguistics.

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*.

Yin Li, Alireza Fathi, and James M Rehg. 2015. Delving into egocentric actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. 2022a. Egocentric video-language pretraining.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. 2022b. Egocentric video-language pretraining. *Neural Information Processing Systems (NeurIPS)*.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*.

Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sudheendra Sinha, Hongxia Song, Huaizu Jin, and Abhinav Gupta. 2015. Learning to recognize activities in videos from unlabeled youtube clips. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4213–4222.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. SimMIM: a simple framework for masked image modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis Langlotz. 2021. Contrastive learning of medical visual representations from paired images and text.

Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. CenterCLIP. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6439–6455, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu. 2023. Deep learning for video-text retrieval: a review. *International Journal of Multimedia Information Retrieval*, 12(1):3.

Linchao Zhu and Yi Yang. 2020. ActBERT: Learning global-local video-text representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.