
Egocentric Video-Language Pretraining

**Kevin Qinghong Lin¹, Alex Jinpeng Wang¹, Mattia Soldan³, Michael Wray²,
Rui Yan¹, Eric Zhongcong Xu¹, Difei Gao¹, Rongcheng Tu⁴,
Wenzhe Zhao⁴, Weijie Kong⁴, Chengfei Cai⁴, Hongfa Wang⁴,
Dima Damen², Bernard Ghanem³, Wei Liu⁴, Mike Zheng Shou^{1*}**

¹Show Lab, National University of Singapore ²University of Bristol

³King Abdullah University of Science and Technology ⁴Tencent Data Platform

Abstract

Video-Language Pretraining (VLP), aiming to learn transferable representation to advance a wide range of video-text downstream tasks, has recently received increasing attention. Dominant works that achieve strong performance rely on large-scale, 3rd-person video-text datasets, such as HowTo100M. In this work, we exploit the recently released Ego4D dataset to pioneer Egocentric VLP along three directions. (i) We create EgoClip, a 1st-person video-text pretraining dataset comprising 3.8M clip-text pairs well-chosen from Ego4D, covering a large variety of human daily activities. (ii) We propose a novel pretraining objective, dubbed as EgoNCE, which adapts video-text contrastive learning to egocentric domain by mining egocentric-aware positive and negative samples. (iii) We introduce EgoMCQ, a development benchmark that is close to EgoClip and hence can support effective validation and fast exploration of our design decisions regarding EgoClip and EgoNCE. Furthermore, we demonstrate strong performance on five egocentric downstream tasks across three datasets: video-text retrieval on EPIC-KITCHENS-100; action recognition on Charades-Ego; and natural language query, moment query, and object state change classification on Ego4D challenge benchmarks. The dataset and code will be available at <https://github.com/showlab/EgoVLP>.

1 Introduction

With the recent interest boom in computer vision and natural language processing, Video-Language Pretraining (VLP) has prevailed, which aims to learn strong and transferable video-language representation for powering a broad spectrum of video-text downstream tasks, such as video-text retrieval [1, 2, 3], video question answering [3, 4, 5, 6], video-captioning [7, 8, 9]. The success of VLP mainly stems from the availability of large-scale open-world video-text datasets [10], which contain a large number of videos sourced from the Web (e.g., YouTube) and pair videos with associated textual information. For instance, HowTo100M [10] collects 134K instructional videos accompanied by noisy narrations yielded from Automatic Speech Recognition (ASR). WebVid-2M [11] scrapes 2.5M descriptive videos with well-formed long captions.

Despite reaching an impressive data scale, videos in the existing video-text pretraining datasets are often of 3rd-person views and might have been edited before posting on the web. Yet, there is a noticeable domain gap between the existing video-text pretraining datasets and 1st-person view videos such as those videos captured by wearable cameras or smart glasses. Egocentric video has received increasing interests from academia (e.g., activity recognition [12], activity anticipation [13], video summarization [14]) and industry (various applications in robotics and augmented reality). But, due

*Corresponding Author.

Dataset	Ego?	Domain	Dur (hrs)	# Clips	# Texts	Example
MSR-VTT [17]	✗	diverse	40	10K	200K	
YouCook2 [18]	✗	cooking	176	14K	14K	
ActivityNet Captions [7]	✗	action	849	100K	100K	
WebVid-2M [11]	✗	diverse	13K	2.5M	2.5M	
HowTo100M [10]	✗	instructional	134K	136M	136M	3rd-person view
Charades-Ego [19]	✓	home	34	30K	30K	
UT-Ego [20]	✓	diverse	37	11K	11K	
Disneyworld [21]	✓	disneyland	42	15K	15K	
EPIC-KITCHENS-100 [22]	✓	kitchen	100	90K	90K	
EgoClip	✓	diverse	2.9K	3.8M	3.8M	1st-person view

Table 1: Comparison of our proposed EgoClip pretraining dataset against the mainstream video-language datasets (top) and egocentric datasets (bottom).

to such a domain gap, directly transferring the existing VLP models to egocentric downstream tasks cannot fully unleash the potential of large-scale pretraining approaches, which we have confirmed in the later experimental section. To bridge this gap, we are motivated to develop Egocentric VLP models, which can greatly benefit various egocentric video downstream applications.

However, existing egocentric video datasets are of small scale and domain-specific, making Egocentric VLP prohibitive. As illustrated in Tab. 1, the formerly largest egocentric video dataset EPIC-KITCHENS-100 [15] focuses on kitchens scenarios and its size is far smaller than the 3rd-person pretraining sets WebVid-2M [11] and HowTo100M [10]. Fortunately, with the recent introduction of the massive-scale egocentric video dataset Ego4D [16], it becomes possible to unlock Egocentric VLP. Ego4D consists of 3,670 hours of videos with manually annotated narrations from 74 worldwide locations, covering a large variety of daily-life scenarios and activities.

In this work, roused by the favorable scale and diversity of Ego4D, we make a significant effort to pave the way for Egocentric VLP with the following steps:

- (i) To address the aforementioned issue of lacking a suitable large-scale egocentric video-language pretraining dataset, we create a video-text pretraining dataset **EgoClip** which contains a total of 3.8M clean 1st-person clip-text pairs selected from Ego4D and cover diverse human daily activities.
- (ii) To make full use of EgoClip for video-text representation learning, we propose a novel video-text contrastive objective **EgoNCE** to address unique challenges in egocentric pretraining datasets.
- (iii) We create a development benchmark i.e. Egocentric Multiple-Choices-Question, dubbed as **EgoMCQ**, which contains 39K questions created from Ego4D and focuses on evaluating video-text alignment. In contrast to other downstream benchmarks, **EgoMCQ** has less discrepancy with EgoClip, powering us to accurately validate and quickly iterate our designs of EgoClip and EgoNCE .
- (iv) We conduct extensive experiments to demonstrate the superiority of Egocentric VLP by transferring our pretrained representation to five egocentric downstream benchmarks. We achieve state-of-the-art performances: 59.4% nDCG on video-text retrieval of EPIC-KITCHENS-100 [15], 32.1% mAP on action recognition of Charades-Ego [19], and significant boosts over three Ego4D challenges ²: natural language query, moment query and object state change classification.

2 Related Work

Video-Language Pretraining. The introduction of large-scale video-text datasets [10, 11] has enabled the emergence of VLP approaches to improve the video-text representation for various video-language tasks [17, 23, 18, 7, 24], such as MIL-NCE [25] proposes to match clips with multiple captions close in temporal to adapt the video-text misalignment of HowTo100M [10]. Dominant VLP methods can be classified into two groups, namely: joint- and dual-encoders. The former combines videos and texts as a single input to the encoder that performs the multimodal fusion. For instance, [26, 27] concatenate videos and texts together before feeding them to a unified transformer. Conversely, methods as [11, 25] exploit dual encoders to independently project the video and text inputs into a common space and minimize the distance between the paired representations. These approaches are to be preferred in retrieval settings as they allow for efficient indexing of a single modality [2, 28]. For example, Frozen [11] employs two separate transformers to encode video and

²<https://ego4d-data.org/Workshop/CVPR22/#challenges>

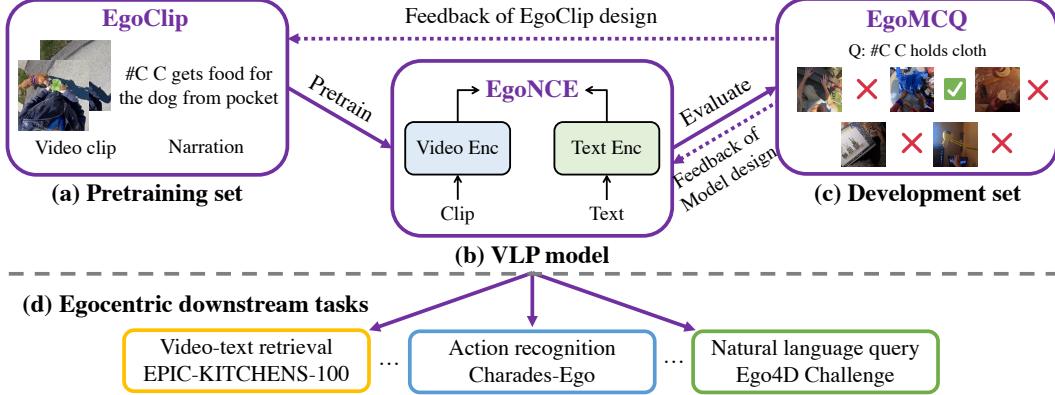


Figure 1: Our Egocentric VLP includes: (a) pretraining set EgoClip; (b) VLP model; and (c) development set EgoMCQ. We use EgoClip to pretrain a VLP model with EgoNCE loss and then evaluate on EgoMCQ. According to the feedback, we iteratively refine our designs of (a) and (b). We then transfer the pretrained model to downstream tasks relevant to the egocentric domain.

text features and align them by video-text InfoNCE [29]. In our work, we adopt the Frozen [11] but extend its InfoNCE to EgoNCE via positive and negative sampling for egocentric-friendly pretraining.

Egocentric Video Datasets. Egocentric videos, collected by participants using wearable cameras, offer a natural perspective of people’s daily activities and raise a range of challenging research topics [14, 20, 12, 13]. Several egocentric video datasets have been developed in decades, e.g., [22, 19, 30]. However, since the collection of egocentric videos is expensive, previous egocentric datasets tend to be small-scale and domain-specific. These limitations hinder 1st-person view research and fail to match the progress of 3rd-person counterparts, such as VLP [25, 26, 11]. Recently, a massive egocentric video dataset Ego4D [16] has been released, which consists of 3,670 hours of videos collected by 931 people from 74 worldwide locations in 9 different countries, where most videos are accompanied by narrations, audio, 3D meshes, and more. Furthermore, Ego4D introduces a suite of new challenging benchmarks (e.g., Natural language query and moment query that we treated as downstream tasks) to fully explore the 1st-person visual experience. With this step-change dataset and benchmarks, Ego4D would lead to a new research surge on egocentric visual perception.

3 EgoClip: An Egocentric Video-Language Pretraining Dataset

Data curation. For our EgoClip dataset, we source data from Ego4D [16], which contains 9,645 untrimmed videos of varying lengths from 5 sec to 7 hrs. From these videos, most are associated with *dense timestamp-level narrations* assigned by two different annotators, describing the camera wearer’s activities and interactions with objects. For example, the narration “#C C puts the scrapper down.” corresponds to video content that occurred at 3.70s, where “#C” refers to the camera-wearer. Notably, narrations in Ego4D are well-aligned with the videos, both temporally and visually. Prior pretraining datasets are characterized by a much greater level of temporal misalignment between the video and text (e.g., HowTo100M [10] narrations are scraped from ASR, yielding sentences misaligned or even unrelated to video content). We first filter Ego4D videos with missing narrations (7.4% of the total video duration) and exclude videos that belong to the validation and test sets of Ego4D benchmark challenge [16] (a further 23.9% of the total video duration). Next, we retain textual annotation from both narrators in EgoClip, allowing us to consider narration diversity when pairing video and text for pretraining purposes. Finally, we adopt several criteria to filter the video and textual narrations, further reducing noise (detailed steps are provided in Supplementary A.1). Overall, this procedure yields 2.9K hours of videos with 3.85 million narrations which covers 2927 hours of video from 129 different scenarios. EgoClip has 21.9 clips per minute with an average clip length of 1.0 seconds and a standard deviation of 0.9 seconds (the longest clip is up to 60s). Additional analyses are included in the Supplementary A.3.

Creation of clip-text pairs. Clip-text pairs are the common data format for VLP, but are usually not present in untrimmed video datasets with only a weak matching between narrations captions and videos. This was first discussed in HowTo100M [10], which pairs subtitles to video clips with

corresponding time intervals to result in noisy pairs. This is not suitable for Ego4D since each narration is annotated with a single timestamp rather than an interval. Thus, we design a *contextual variable-length clip pairing strategy*. Formally, narrations per video in Ego4D are organized as a sequence of sentences $\{\mathcal{T}_0, \dots, \mathcal{T}_n\}$ with exact timestamps $\{t_0, \dots, t_n\}$, indicating an event i described by \mathcal{T}_i happened in the moment t_i . For a narration \mathcal{T}_i with timestamp t_i , we pair a clip \mathcal{V}_i with following start and end timepoints:

$$[t_i^{start}, t_i^{end}] = [t_i - \beta_i/2\alpha, t_i + \beta_i/2\alpha], \quad (1)$$

which represents a window centered around the timestamp t_i with temporal duration equal to β_i/α . β_i is an adjustable parameter equal to the average temporal distance between pairs of consecutive narrations, i.e., $\sum_{j=0}^{n-1} (t_{j+1} - t_j)/n$. We compute β_i on a per video basis. Conversely, α is a scale factor computed as the average of all β_i across all videos in the EgoClip ($\alpha = 4.9$ seconds). Intuitively, Eq. 1 is derived from three observations: **(i)** Centering t_i helps to involve prior information about the event i ; **(ii)** β_i measures the clip duration according to its scenario, such as longer clips watching television (352.9 seconds) v.s. shorter clips harvesting crops (0.9 seconds); **(iii)** α controls the context granularity of clips (e.g., a large α pays more attention to rapid, atomic actions). We ablate these design choices in our experimental section.

4 Video-Language Pretraining Model

To efficiently transfer video-language representation to egocentric downstream tasks (e.g., video-text retrieval on EPIC-KITCHENS-100 [22]), We prefer the dual-encoder (discussed in Sec. 2) as our VLP model architecture. In particular, we emphasize devising a general pretraining objective EgoNCE to adapt existing VLP model to the egocentric environment (e.g., EgoClip).

4.1 Architecture: Dual-encoder Pipeline

We choose Frozen [11] as our pretraining architecture. Frozen [11] design encompasses an elegant and simple dual encoder strategy (one per modality) which has favorable characteristics (e.g., indexability and efficiency [2, 28]). Note that this allows us to use our pretrained network in single-modality tasks (e.g., video-only tasks). In practice, the video encoder is a TimeSformer [31] architecture while the text encoder builds upon DistillBERT [32]. However, our approach is not limited to the encoder’s design (e.g., the video backbone can be replaced by SlowFast [33]). In the rest of the paper we adopt this notation: $(\mathcal{V}_i, \mathcal{T}_i)$ represent the video-test input to the model, while \mathbf{v}_i and \mathbf{t}_i are used to identify the video and text embedding.

4.2 EgoNCE: An Egocentric-friendly Pretraining Objective

A common pretraining objective for dual-encoder VLP is InfoNCE [29], where the matching visual-text pairs in the batch are treated as positives while all other pairwise combinations in the batch are regarded as negatives. Formally, within a batch $\mathcal{B} = \{1, \dots, N\}$, InfoNCE is computed by the sum of video-to-text loss \mathcal{L}_{v2t} and text-to-video loss \mathcal{L}_{t2v} . For simplicity, we only formulate \mathcal{L}_{v2t} whereas \mathcal{L}_{t2v} is defined in a symmetry way.

$$\mathcal{L}_{v2t} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\mathbf{v}_i^T \mathbf{t}_i / \tau)}{\sum_{j \in \mathcal{B}} \exp(\mathbf{v}_i^T \mathbf{t}_j / \tau)}, \quad (2)$$

where i -th video embedding \mathbf{v}_i and j -th text embedding \mathbf{t}_j are L2 normalized features, and τ is a temperature factor.

However, this simple objective performs not well on large-scale video-text datasets like HowTo100M [10] due to the serious misalignment between the two modalities. Therefore, [34] proposes MIL-NCE which treats temporal nearest captions as positive samples.

In this work, our 1st-person human daily activity dataset, i.e. EgoClip, presents two unique challenges compared to the existing 3rd-person view video-text datasets: **Challenge (i):** The *same action* often occurs in *different scenarios* (e.g., “unlock the phone” could happen when “lying in bed” or “walking outdoors”). **Challenge (ii):** Often, *different actions* appearing in the *same scenario* tend to have indistinguishable visual differences (e.g., when “working in front of the laptop”, “typing on the keyboard” or “moving the mouse” could have similar feature representations).

Evaluation on the text-video retrieval task is unreliable due to duplications

Retrieval result: Top clips are not GT but shall be considered as correct.

Text query:	#C C closes the refrigerator.	Top1:		Top2:		Top3:		...	TopN (GT):	
	#C C closes the refrigerator.				#C C closes the fridge				#C C closes the lower part of the fridge	#C C closes the refrigerator.

EgoMCQ	Inter-video					Intra-video				
Text query	#C C picks the silicone sealant					#C C carries paint bucket down the ladder				
Select the correct video clip from 5 candidates										
Answer with GT	#C C places the camping seat down	#C C holds the power drill with both hands.	#C C picks the silicone sealant	#C C takes a stone	#C C cuts the green bean into pieces	#C C holds paintbrush with both hands	#C C turns paintbrush in his left hand	#C C shifts paintbrush to right hand	#C C drops paintbrush on paint bucket	#C C carries paint bucket down the ladder

Figure 2: Design of Egocentric VLP development set. **Top:** Illustration of why the task of text-video retrieval is not suitable; **Bottom:** Two settings of EgoMCQ. **Left-bottom:** The “inter-video” setting, each question contains 5 clips from different videos. **Right-bottom:** The “intra-video” setting, each question contains 5 contiguous clips (after dedup) from the same video, making it more challenging. To overcome these two unique challenges, we propose a novel EgoNCE training objective which takes into account two simple yet efficient sampling strategies based on the vanilla InfoNCE.

Action-aware Positive Sampling. In this work, we make a reasonable assumption that the critical elements in linking visual actions to textual narrations are verbs and objects mentioned in the narrations (e.g. “drinking coffee”, “opening fridge”). Following this assumption, we can devise a clever method to address challenge (i). Specifically, for each narration, we identify its nouns and verbs and merge synonym words based on the Ego4D taxonomy dictionary [16], a thesaurus records meaningful nouns/verbs in Ego4D narrations. Then, batch samples that shared at least one noun and at least one verb are treated as positive samples. At last, for the sample i , we select its positive samples set within batch \mathcal{B} as $\mathcal{P}_i = \{j \in \mathcal{B} \mid \text{noun}(j) \cap \text{noun}(i) \neq \emptyset, \text{verb}(j) \cap \text{verb}(i) \neq \emptyset\}$.

Scene-aware Negative Sampling. To address challenge (ii), we consider different actions in the same scenario as hard negative samples. Specifically, for each video clip i , we sample an adjacent clip $i' \in \mathcal{N}(i)$, which is close to i in time within the same video. We augment the original batch \mathcal{B} with such hard negative samples and each sample i in \mathcal{B} has its negative counterparts i' . Hence the batch is updated as $\tilde{\mathcal{B}} = \underbrace{\{1, 2, \dots, N\}}_{\mathcal{B}}, \underbrace{\{1', 2', \dots, N'\}}_{\mathcal{N}(\mathcal{B})}$.

With these two sampling strategies, our new pretraining objective **EgoNCE** can be formulated as:

$$\mathcal{L}_{\text{v2t}}^{\text{ego}} = \frac{1}{|\tilde{\mathcal{B}}|} \sum_{i \in \tilde{\mathcal{B}}} \log \frac{\sum_{k \in \mathcal{P}_i} \exp(\mathbf{v}_i^T \mathbf{t}_k / \tau)}{\sum_{j \in \mathcal{B}} (\exp(\mathbf{v}_i^T \mathbf{t}_j / \tau) + \exp(\mathbf{v}_i^T \mathbf{t}_{j'} / \tau))}. \quad (3)$$

Here the item in purple corresponds to our proposed action-aware positive samples and blue corresponds to our proposed scene-aware negative samples. EgoNCE provides a general extension to adapt the existing VLP models for video-text pretraining datasets in the egocentric domain.

5 EgoMCQ: A Benchmark for Egocentric VLP Development

The need for a development benchmark. We find that most egocentric benchmarks are domain-specific and focus on single-modality tasks (see Tab. 1). However, our purpose is to exploit Ego4D’s diversity to learn rich video-text representations. Hence, to validate our design choices of the pretraining dataset (e.g., EgoClip), and model (e.g., EgoNCE), it is essential to measure performance on a benchmark highly aligned with the pretraining task. Therefore, we propose EgoMCQ, a new egocentric benchmark for reliable and fast developments of Egocentric VLP.

Data source. We start from the Ego4D data excluded from constructing the EgoClip, which mainly covers the validation set of the Ego4D challenge benchmarks. Additionally, to assure that the scene is not visible during pretraining, we manually remove videos that share multiple views with the videos in EgoClip. To ensure diversity, we randomly select one annotator’s narration for each video. We follow the same clip pairing strategy as Eq. 1 to be consistent with the data format of EgoClip preventing a domain gap.

Benchmarking task design. To determine the task for development, we first consider video-text retrieval since it highly aligns with the VLP pretraining objective. However, as depicted in the top half of Fig. 2, for an action (e.g., close the refrigerator), there are substantial duplicate/semantically similar captions in Ego4D. This can cause issues in retrieval evaluation [35] causing model training to become unreliable. A straightforward approach to prevent this is deduplication (dedup), but it is challenging to devise a dedup criterion and perform well in the retrieval settings of a “one-to-whole validation set”. Therefore, we select the *Multiple-Choice Questions (MCQ)* task for development since repetitions are highly unlikely given a small number of answers.

Grouping strategies. To set up the MCQ task, a naive construction randomly groups five video clips to form options for a question. But we find randomly grouping is not challenging since options are highly likely to come from different videos and vary widely in content. We redefine this basic setting as “**inter-video**” and ensure that the five clips originate from different videos, aiming to distinguish instances from different scenarios (the left side of Fig. 2). Furthermore, we propose a more challenging setting “**intra-video**” by grouping five continuous clips together (with dedup). This setting is regarded as a specific form of video-text localization focused on fine-grained context clues, such as hand interaction (the right side of Fig. 2). Dedup is performed within five options for each question for reliable assessment (see Supp B.1) and we adopt accuracy as the EgoMCQ metric.

Statistics. We finalize 39K questions covering 198K narrations with 468 hours of video, where the “inter-video” has 24K questions covering 290.3 hours of videos. And the “intra-video” has 15K questions and covers 178.3 hours of videos. The average duration among the five options is 34.2 seconds (More statistics of EgoMCQ are shown in Supplementary B.2.).

6 Experiments

We assess our Egocentric VLP along two directions: **(i)** We conduct extensive analysis to explore key components of Egocentric VLP (e.g., EgoClip, EgoNCE, and EgoMCQ); **(ii)** We transfer our pretrained model to various downstream tasks to validate the quality of our video-text representation.

6.1 Benchmarks and Settings

We evaluate our VLP model on five egocentric benchmarks, spanning video-text tasks and pure video tasks, across three different datasets. We briefly describe each task below.

Multi-Instance Retrieval of EPIC-KITCHENS-100. This task is modelled as a text-to-video retrieval which considers the semantic overlap between different videos narrations, where multiple videos may correspond to the same narration. The training set contains 67.2K clips and validation set contains 9.7K clips. The evaluation metrics are mean Average Precision (mAP) and normalised Discounted Cumulative Gain (nDCG).

Natural Language Query of Ego4D Challenges. The Natural Language Query task is modelled as a natural language grounding problem [1, 36, 37]. Given a language query and a video, the task aims at localizing the temporal interval within the video, in which the answer is deducible. The training set contains 11.3K queries annotated from 1K clips for this task, while the validation contains 3.9K queries collected from 0.3K clips. The evaluation metric is Recall@ K for IoU=θ (R@ K -IoU=θ) [1] where θ is a threshold. We evaluate for $K \in \{1, 5\}$ and $\theta \in \{0.3, 0.5\}$.

Action Recognition of Charades-Ego. This dataset has 64K instances spanning 1st-person and third-person views, and covering 157 activity categories for training. We train and evaluate only on the 1st-person videos. The validation set contains 847 videos for classification and each video belongs to multiple classes. The evaluation metric is mAP.

Moment Query of Ego4D Challenges. The Moment Query task is a video-only task modelled as Temporal Action Localization [12]. Given a particular high-level activity category, the task solution

Clip creation strategy	Clip's length (s) Avg \pm Std	EgoMCQ Acc (%)		Zero-shot T \leftrightarrow V Retrieval [22]	
		Inter-video	Intra-video	mAP (avg)	nDCG (avg)
(a) $[t_i, t_i+\alpha]$	5.0 ± 0.0	87.66	39.72	19.6	12.3
(b) $[t_i-\alpha/2, t_i+\alpha/2]$	5.0 ± 0.0	89.23	41.68	20.6	13.7
(c) $[t_{i-1}, t_{i+1}]$	10.0 ± 38.2	88.13	40.62	20.6	13.7
(d) $[t_i-\beta_i/2, t_i+\beta_i/2]$	4.9 ± 4.7	<u>89.74</u>	44.82	21.1	14.5
(e) $[t_i-\beta_i/4, t_i+\beta_i/4]$	2.4 ± 2.4	90.23	49.67	<u>21.9</u>	<u>15.3</u>
(f) $[t_i-\beta_i/2\alpha, t_i+\beta_i/2\alpha]$	1.0 ± 0.9	89.36	51.51	<u>22.1</u>	<u>15.5</u>

Table 2: Results on our development set EgoMCQ and video-text retrieval on EPIC-KITCHENS-100 when using different strategies in the creation of EgoClip, where t_i , α , β_i are defined in Eq. 1. In all experiments, we bold the **best results** and underline the second best results.

consist of retrieving all the possible temporal windows where the activity occurs. The training set contains 13.6K instance from 1.5K clips, while the validation set contains 4.3K instances from 0.5K clips. The evaluation metrics are mAP and R@ K -IoU= θ for $K \in \{1, 5\}$ and $\theta \in \{0.3, 0.5, 0.7\}$.

Object State Change Classification (OSCC) of Ego4D Challenges. This OSCC task is modelled as an (N+1)-way classification aiming to identify an object’s state change in a given video. The training and val. sets contain 41K and 28K clips, respectively. The evaluation metric is accuracy.

Implementation details. Our codebase is based on the official Frozen³ one and retains the same settings unless specified. During pretraining, we sample 4 frames for each clip, and use the Adam optimizer [38] with a learning rate of 3×10^{-5} . To select the best method we pretrain our architecture for 10 epochs and use the best performing model on the EgoMCQ benchmark. Pretraining takes two days on 32 A100 GPUs (1,536 GPU hrs).

6.2 Ablation Studies

Ablation of the strategy used when creating EgoClip. We validate our proposed strategies i.e. Eq.1 here by comparing the following variants: (a) fixed length α , start at timestamp; (b) fixed length α , center at timestamp; (c) variable clip, start and end by adjacent timestamps; (d) our proposed strategy, scaled by 2; (e) our proposed strategy, scaled by 4; (f) our proposed strategy.

We consider a good pretraining dataset creation strategy should satisfy the following: (1) the VLP model trained on EgoClip should be able to well distinguish instances in EgoMCQ with the same data format; (2) the VLP model pretrained on EgoClip with the specific clip creation strategy should perform well in public downstream tasks (e.g., video-text retrieval on [22], zero-shot for efficiency).

We draw several conclusions from Tab. 2: (i) The performance of EgoMCQ is well aligned with the zero-shot result on EPIC-KITCHENS-100, especially minor gain on downstream but noticeable on EgoMCQ, which means EgoMCQ provides valid feedback and is suitable as a development set. (ii) Under the same clip length α , (b) surpassing (a) proves that centering at timestamp includes prior information is helpful. (iii) Variable-length clips make a big difference, as shown in (c) and (d).

Notably, with our designed β_i , (d) outperforms (b) with a similar average clip length, which validates our key idea of “contextual varied clip length”. (iv) Based on (d), (e), and (f), we found a proper scale factor great than 1 is preferred, which helps to focus on a large of instantaneous actions densely labeled by Ego4D [16]. These ablation studies demonstrate the effectiveness of our proposed EgoClip creation strategy and EgoMCQ for development.

Effect of EgoNCE. In this section, we evaluate the effect of the proposed sampling strategies for the EgoNCE objective (Eq. 3) on EgoMCQ and compare against a vanilla InfoNCE loss (Eq. 2). We ablate several configurations for positive and negative sampling strategies. The sampling strategy for positive pairs exploits language cues, while negative pairs rely on temporal, visual cues. Given a text-video pair, we regard other text-video pairs as positive if the textual narrations: (a) share at least one noun, (b) share

Variants	Accuracy (%)	
	Intra-video	Inter-video
InfoNCE	89.4	51.5
(a) w/ Pos, noun	82.9 (6.5 \downarrow)	42.3 (9.2 \downarrow)
(b) w/ Pos, verb	86.9 (2.5 \downarrow)	50.5 (1.0 \downarrow)
(c) w/ Pos, noun & verb	<u>89.7</u> (0.4 \uparrow)	53.6 (2.1 \uparrow)
(d) w/ Neg, random	88.3 (1.1 \downarrow)	49.9 (1.6 \downarrow)
(e) w/ Neg, within video	<u>89.7</u> (0.3 \uparrow)	53.0 (1.5 \uparrow)
(f) w/ Neg, within 1 min	89.5 (0.2 \uparrow)	<u>54.5</u> (3.0 \uparrow)
(g) w/ Pos & Neg, EgoNCE	90.6 (1.3 \uparrow)	57.2 (5.7 \uparrow)

Table 3: Pretraining sampling strategy ablation. We evaluate accuracy performance on our development benchmark EgoMCQ.

³<https://github.com/m-bain/frozen-in-time>

Methods	Vis Enc Input	# Frames	Vis-text PT	mAP (%)			nDCG (%)		
				V→T	T→V	Avg	V→T	T→V	Avg
Random	-	-	-	5.7	5.6	5.7	10.8	10.9	10.9.
MI-MM	S3D [39]	32	HowTo100M	34.8	23.6	29.2	47.1	42.4	44.7
MME [40]	TBN † [15]	25	-	43.0	34.0	38.5	50.1	46.9	48.5
JPoSE [40]	TBN † [15]	25	-	49.9	38.1	44.0	55.5	51.6	53.5
Frozen	Raw Videos	4	HowTo100M	38.8	29.7	34.2	50.5	48.3	49.4
Frozen	Raw Videos	4		39.2	30.1	34.7	50.7	48.7	49.7
Frozen	Raw Videos	4		41.2	31.6	36.4	52.7	50.2	51.4
Frozen	Raw Videos	4		44.5	34.7	39.6	55.7	52.9	54.3
Frozen+EgoNCE	Raw Videos	4	EgoClip	45.1	35.3	40.2	56.2	53.5	54.8
Frozen	Raw Videos	16	CC3M+WebVid2M	45.8	36.0	40.9	57.2	54.3	55.8
Frozen+EgoNCE	Raw Videos	16	EgoClip	49.9	40.5	45.0	60.9	57.9	59.4
Frozen	Raw Videos	4	HowTo100M.	6.8	6.3	6.5	11.6	12.8	12.2
Frozen	Raw Videos	4	CC3M+WebVid2M.	8.6	7.4	8.0	14.5	14.6	14.5
Frozen	Raw Videos	4	EgoClip.	17.9	13.1	15.5	23.0	21.2	22.1
Frozen+EgoNCE	Raw Videos	4	EgoClip	19.4	13.9	16.6	24.1	22.0	23.1

Table 4: Performance of the EPIC-KITCHENS-100 Multi-Instance Retrieval. Note that TBN † feature [15] are a combination of three modalities: RGB, Flow and Audio. Conversely, our approach only relies on RGB input. The grey highlighted rows correspond to **zero-shot evaluation**.

at least one verb, (c) share at least a verb-noun pair. Conversely, we define the following heuristics for negative sampling: (d) a random text-video pair from EgoClip, (e) a text-video pair from the same video, (f) a text-video pair within 1 minute from the given video-text pair annotation timestamps. Tab. 3 shows that only using verbs (a) or nouns (b) for positive selection degrades the accuracy performance with respect to naive InfoNCE. However, we successfully push the performance beyond the baseline results when considering both verbs and nouns jointly (c). Moreover, we notice that only selecting negatives within the same video leads to better performance. In particular, we obtain the best performance for temporally “hard negatives” (f). Finally, we pick the optimal settings from positive and negative samplings and combine them together for (g) EgoNCE and reach the best results.

6.3 Comparisons with State-of-the-arts

Multi-Instance Retrieval. In Tab. 4, we report both zero-shot and fine-tuning evaluation results. In the zero-shot setting, pretraining with EgoClip (3.8M), despite being smaller in scale, still outperforms CC3M+WebVid-2M (5.5M) and HowTo100M (136M), validating the unique benefit of pretraining on egocentric data. When fine-tuned with 4 frames (rows 5-9), EgoClip pretraining maintains a margin over the best baseline CC3M+WebVid-2M, further verifying the viewpoint domain gap within fine-tuning. Lastly, we increase the sample frames of our finalized model as well as best competitor CC3M+WebVid-2M pretraining to 16 (rows 10-11). As expected, performance gains accompany the frame increase. We deem that notable benefits come from better temporal modeling for frequent action interactions in the 1st-person view. Overall, our pretraining model outperforms the best baseline (JPoSE) by 1.0 mAP and 5.9% nDCG while requiring fewer frames and input modalities.

Action Recognition. We conduct action recognition on Charades-Ego, where categories are short phrases like “Holding some clothes”. Thus this task can be solved as a text-to-video retrieval by leveraging the text representation. We present the result in Tab. 5 under zero-shot and fine-tuning settings. In zero-shot settings, our model outperforms two supervised baselines, which validate the stronger generalization of jointly video-text feature. After fine-tuning (rows 5-9), our model surpasses

Methods	Vis Enc	# Frames	Vis-Text PT	Train / FT Data	mAP (%)
Actor [41]	ResNet-152	25	-	Charades-Ego (1st + 3rd)	20.0
SSDA [42]	I3D	32	-	Charades-Ego (1st + 3rd)	23.1
I3D [42]	I3D	32	-	Charades-Ego (1st)	25.8
Ego-Exo [43]	SlowFast (Res-101)	32	-	Charades-Ego (1st)	30.1
Frozen	TimeSformer	16	-	Charades-Ego (1st)	28.8
Frozen	TimeSformer	16	HowTo100M	Charades-Ego (1st)	28.3
Frozen	TimeSformer	16	CC3M+WebVid2M	Charades-Ego (1st)	30.9
Frozen	TimeSformer	16	EgoClip	Charades-Ego (1st)	31.2
Frozen+EgoNCE	TimeSformer	16	EgoClip	Charades-Ego (1st)	32.1
Frozen	TimeSformer	16	HowTo100M	-	9.2
Frozen	TimeSformer	16	CC3M+WebVid2M	-	20.9
Frozen	TimeSformer	16	EgoClip	-	23.6
Frozen+EgoNCE	TimeSformer	16	EgoClip	-	25.0

Table 5: Performance of the action recognition on Charades-Ego dataset (First-person test set). The grey highlighted rows correspond to **zero-shot evaluation**.

Methods	Video Pre-extracted Features		IoU=0.3		IoU=0.5		IoU=0.7		mAP (%) @ IoU			
	Vis Enc	Vis-text PT	R@1	R@5	R@1	R@5	R@1	R@5	0.1	0.3	0.5	Avg
VSGN [45]	SlowFast	-	33.45	58.43	25.16	46.18	15.36	25.81	9.10	5.76	3.41	6.03
VSGN [45]	Frozen	HowTo100M	31.40	52.61	22.28	41.29	13.41	23.21	9.83	6.72	3.84	6.72
VSGN [45]	Frozen	CC3M+WebVid2M	32.08	56.40	23.46	43.81	13.73	23.77	9.83	6.40	3.86	6.58
VSGN [45]	Frozen	EgoClip	40.06	63.71	29.59	48.32	17.41	26.33	15.90	10.54	6.19	10.69
VSGN [45]	Frozen+EgoNCE	EgoClip	40.43	65.67	30.14	51.98	19.06	29.77	16.63	11.45	6.57	11.39

Table 7: Recall and mAP metrics for several IoU on the Moment Query task’s val. set.

all VLP counterparts and improves over the state-of-the-art video classifier Ego-Exo by 2.0% with fewer sampled video frames, which shows the superior advantage of joint video-text representations.

Natural Language Query. We report validation results on Tab. 6. We adopt the same baselines as introduced in [16], namely: 2DTAN [24] and VSLNet [44], and substitute the SlowFast-BERT features with our video and language representations. We observe a large boost in performance offered by our pretrained model for all metrics. Notably, we improve R@1 for IoU=0.3 from 5.45 to 10.84, despite our video branch not being pre-trained on Kinetics400. Besides, we significantly surpass VLP pretrained on CC3M+WebVid-2M and HowTo100M. We believe this increase is due to the egocentric data availability and the video-text interaction learned from large-scale pretraining. See Supplementary D.6 for test set results.

Moment Query. This task investigates the quality of video-only features. We extract video features and provide them as input to the VSGN model [45]. We report the validation results in Tab. 7. We find that our features achieves the best performance over SlowFast features with an increase of 4.66% in Avg mAP. Moreover, we maintain better performance with respect to 3rd-person large-scale pretraining datasets. This demonstrates that the 1st-person VLP model also learns competitive video representations. See the Supplementary D.5 for test set results.

Object State Change Classification. We report the validation results on Tab. 8. Once again, our model achieves the best performance of all baselines, 2.4% than CC3M+WebVid-2M counterparts, which indicates our visual representations are able to focus on the fine-grained clues related to changes.

Summary of EgoNCE. From the above experimental results, Frozen pretrained on EgoClip with EgoNCE objective brings a consistent improvements over the InfoNCE in all downstream tasks, which comprehensively demonstrates the effect of EgoNCE, and the decision from EgoMCQ.

Methods	Video-text Pre-extracted Features		IoU=0.3		IoU=0.5	
	Vis-text Enc	Vis-text PT	R@1	R@5	R@1	R@5
2D-TAN [24]	SlowFast+BERT	-	-	-	5.04	12.89
VSLNet [44]	SlowFast+BERT	-	-	-	5.45	10.74
VSLNet [44]	Frozen	HowTo100M	3.95	8.72	2.01	4.62
VSLNet [44]	Frozen	CC3M+WebVid2M	5.06	10.30	2.71	6.69
VSLNet [44]	Frozen	EgoClip	10.53	17.94	5.96	11.85
VSLNet [44]	Frozen+EgoNCE	EgoClip	10.84	18.84	6.81	13.45

Table 6: Recall for several IoU on the NLQ task’s val. set.

Methods	Vis-Text PT	Acc. (%)
Always Positive	-	48.1
Bi-d LSTM [46]	ImageNet	65.3
I3D ResNet-50 [47]	-	68.7
Frozen	-	70.3
Frozen	HowTo100M	71.7
Frozen	CC3M+WebVid2M	71.5
Frozen	EgoClip	73.4
Frozen+EgoNCE	EgoClip	73.9

Table 8: Accuracy metric on the Object State Change Classification task’s val set.

7 Conclusion, Limitations, and Societal Impacts.

To the best of our knowledge, this work is the pioneering work to unlock Egocentric VLP. **(i)** We devise a principled data curation and create EgoClip, an egocentric large-scale text-video pretraining dataset with 3.8M clip-text pairs well-chosen from Ego4D. **(ii)** We exploit the particular characteristics of egocentric videos and devise EgoNCE with meaningful sampling strategies for effective egocentric pretraining. **(iii)** We create EgoMCQ, an egocentric video-language benchmark close to the pretraining set to support efficient exploration and development of EgoClip and EgoNCE. Finally, we further demonstrate the strong representation of our egocentric pretraining on five tasks across three datasets. We believe our EgoClip, EgoMCQ and EgoNCE would greatly benefit the egocentric video community, laying down strong foundation for this new scientific topic of egocentric VLP.

Limitations. Our pretraining approach does not take into account the long-term temporal dependencies in long Ego4D videos. We leave it for future work.

Societal impact. We acknowledge that the intensive electricity consumption related to our research has a significant carbon footprint that is negative for the environment. The research community should favor more efficient strategies for computation-intensive applications to limit this issue.

References

- [1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337*, 2018.
- [2] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*, 2019.
- [3] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *IJCV*, 123(1):94–120, 2017.
- [4] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021.
- [5] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, pages 471–487, 2018.
- [6] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755, 2020.
- [7] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.
- [8] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.
- [9] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, pages 8739–8748, 2018.
- [10] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019.
- [11] Max Bain, Arsha Nagrani, Gülcin Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021.
- [12] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.
- [13] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *CVPR*, pages 5343–5352, 2018.
- [14] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *MM*, pages 533–542, 2002.
- [15] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, pages 5492–5501, 2019.
- [16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*, 2021.
- [17] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.
- [18] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.

- [19] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.
- [20] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, pages 1346–1353. IEEE, 2012.
- [21] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *CVPR*, pages 1226–1233. IEEE, 2012.
- [22] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 130(1):33–55, 2022.
- [23] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yuetong Zhuang. Video question answering via gradually refined attention over appearance and motion. In *MM*, pages 1645–1653, 2017.
- [24] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, volume 34, pages 12870–12877, 2020.
- [25] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020.
- [26] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341, 2021.
- [27] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, October 2019.
- [28] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, pages 9826–9836, 2021.
- [29] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.
- [30] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015.
- [31] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095*, 2(3):4, 2021.
- [32] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [33] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.
- [34] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020.
- [35] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *CVPR*, 2021.
- [36] Gao Jiyang, Sun Chen, Yang Zhenheng, Nevatia, Ram. TALL: Temporal Activity Localization via Language Query. In *ICCV*, 2017.
- [37] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. In *ICCV*, pages 3224–3234, 2021.
- [38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [39] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018.

- [40] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *ICCV*, pages 450–459, 2019.
- [41] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, pages 7396–7404, 2018.
- [42] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *WACV*, pages 1717–1726, 2020.
- [43] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*, pages 6943–6953, 2021.
- [44] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020.
- [45] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *ICCV*, pages 13658–13667, 2021.
- [46] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional lstm networks for improved phoneme classification and recognition. In *ICANN*, pages 799–804. Springer, 2005.
- [47] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.
- [48] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Appendix

We present the following items in the supplemental material:

- A. Construction details and statistics of EgoClip pretraining dataset in Sec.A.
- B. Construction details and statistics of EgoMCQ benchmark in Sec.B.
- C. Technical details of our VLP model in Sec.C.
- D. Additional experimental details and results in Sec.D.

A Construction details and statistics of EgoClip pretraining dataset

A.1 Data filtering

After we source video-text data for EgoClip, we adopt the following criteria to further reduce noise:

- (i) We select double-sized stereo videos (1.3% videos dur) and keep half per video for a normal size.
- (ii) We discard videos with an aspect ratio greater than 2 (0.4% videos dur).
- (iii) We filter narrations with unsure tags (4.0% texts) e.g. “#C C washes #unsure in sink”.
- (iv) We remove narrations less than 3 words (0.9% texts), since such narrations generally cannot be deduced from the video, e.g., “#C C speaks”, “#C C looks”.

A.2 Data compression

The Ego4D videos are untrimmed, which tend to be very long (average 24 mins and max to 7 hrs) and have large resolution (e.g., 1920×1080 , 1440×1080), so it is impossible to adopt untrimmed videos as model input due to heavy data loading. Therefore we propose to compress them:

- (i) We first resize all videos with short size 256.
- (ii) Chunk each resized video into several segments, which are up to 10 min in length.

During pretraining, given the start and end time points of a clip, we only load the segment that this clip belongs to, rather than the whole video. To this end, we are able to perform efficient end-to-end pretraining with raw RGB videos as model input. One epoch of pretraining 3.8M video-text pairs costs 6 hrs on 32 V100 GPUs (192 GPU hrs).

A.3 Data analysis

Geographic diversity. We present the distribution of EgoClip clips source in Fig 3, which covers worldwide 13 institutions from 9 different countries [16], including: Europe (UK, Italy); Asia (India, Japan, Singapore, Kingdom of Saudi Arabia); America (USA, Colombia); Africa (Rwanda). Therefore, our created pretraining dataset inherited the good geographic as well as participants diversities of Ego4D (More details can be found in “Supp. C. Demographics” in Ego4D paper [16]).

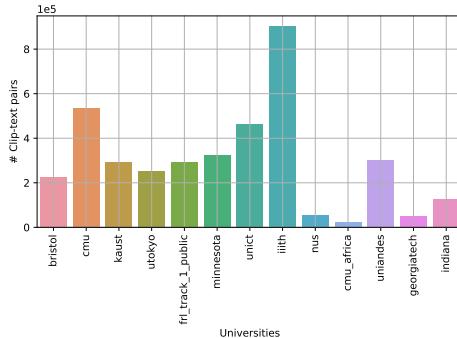


Figure 3: Institution distribution of EgoClip



Figure 4: Scenario distribution of EgoClip



Figure 5: Scenario distribution of EgoMCQ

Scenario diversity. We have statistics the scenario distribution of EgoClip in Fig 4, which covers 129 human daily scenarios e.g., household (cooking, cleaning), outdoor (shopping, hiking), workplace (at desk, on a laptop), leisure (playing board games), etc. Notably, this distribution is long tailed, where the largest scenario “Crafting/knitting/sewing/drawing/painting” includes 622K (11.1%) and the smallest scenario “Hair and Makeup stylist” contains 35 instances.

Clip analysis. We present the statistics on the created clips in EgoClip. Fig 6 (a) shows the distribution of clip frequency over the 2.9K pretraining set videos (For each video, we calculate two frequencies from two annotators respectively). The varying clip frequencies are mainly dependent on manual narrations that are annotated based on the video scenarios and activities. There have average 13.4 clips per minute of video, maximize to 175.8 narrations / minute and minimize to 0.06 narrations / minute. Our clip creation strategy Eq. (1) takes this characteristic into account by estimating clip length based on the frequency of the video that the clip belongs. Fig 6 (b) displays the distribution of clip duration. The average duration is 0.98 seconds with a standard deviation of 0.95 seconds, and 69.5% of clips are less than 1.0 seconds in length, due to the massive atomic instantaneous actions densely labeled by Ego4D. Besides, the clip might be max to 65.36 seconds, which corresponding to the scenario that “a people walking in a forest”.

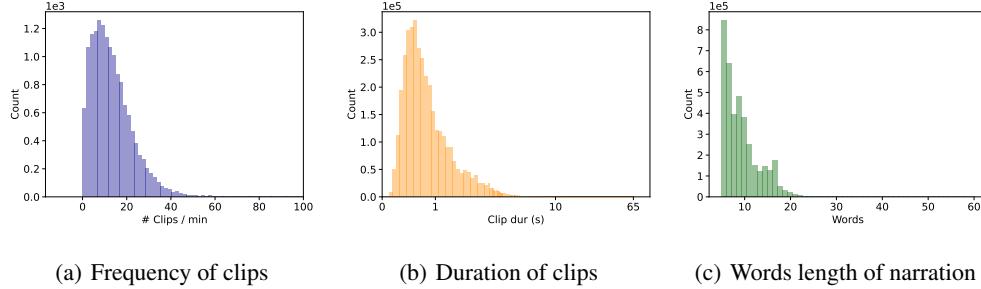


Figure 6: Clip and narration distribution of EgoClip

Narration analysis. In Fig 6 (c), we present the distribution of narration words length. The average words length of EgoClip narration is 9.39. Notably, the EgoClip narrations cover 116 verbs and 555 nouns, where we merge the semantically synonyms words, e.g., the nouns of “handkerchief”, “napkin”, “serviette”, “tissue”, “wipe” both belong to “napkin”. Each narration of EgoClip have 1.84 nouns and 0.87 verbs on average.

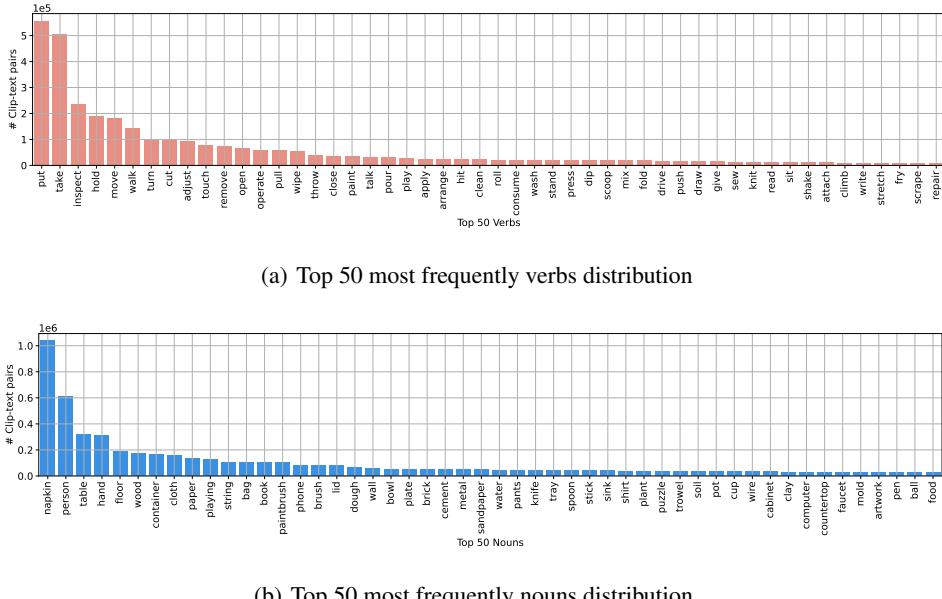


Figure 7: Verbs and nouns distributions of EgoClip’s narrations

We further display the distribution of the top 50 most frequently verbs and nouns of EgoClip in Fig 7. The most common nouns is “napkin”, which appeared in 1.0M (27.06%) clips.

Visualizations. In Fig 8, we visualize some clip-text pairs created by our strategy.



(a) #C C ties the vegetable with a band.



(b) #C C moves the right hand.



(c) #C C picks the chopsticks.



(d) #C C cuts the apple with a knife.



(e) #C C draws on a book.

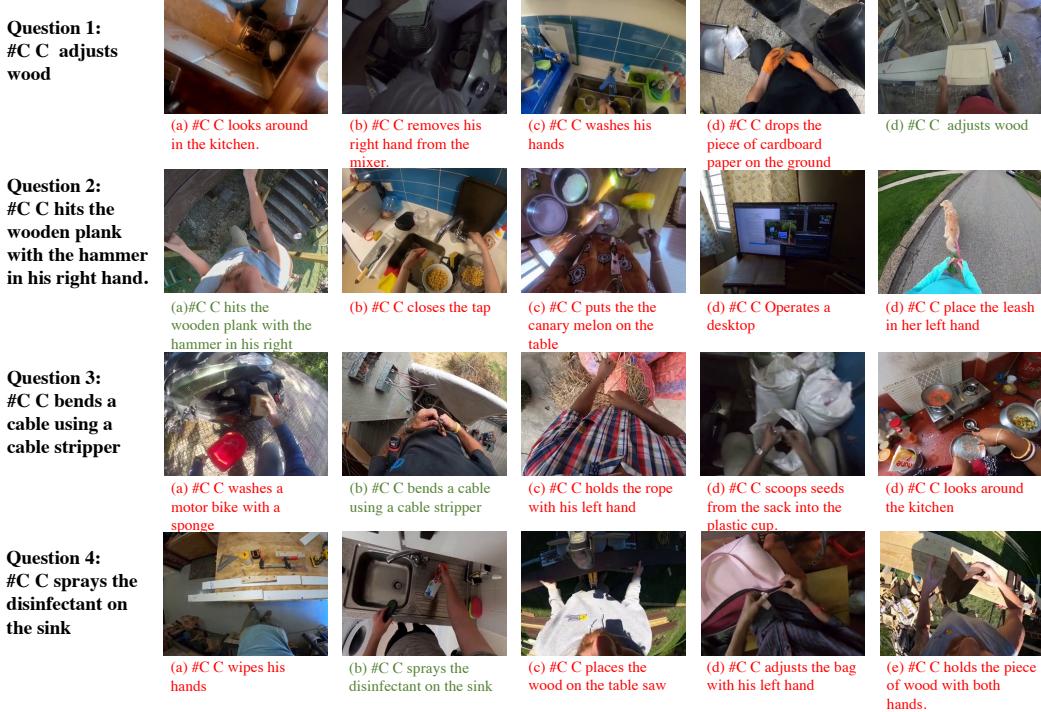


(f) #C C stretches his left hand.

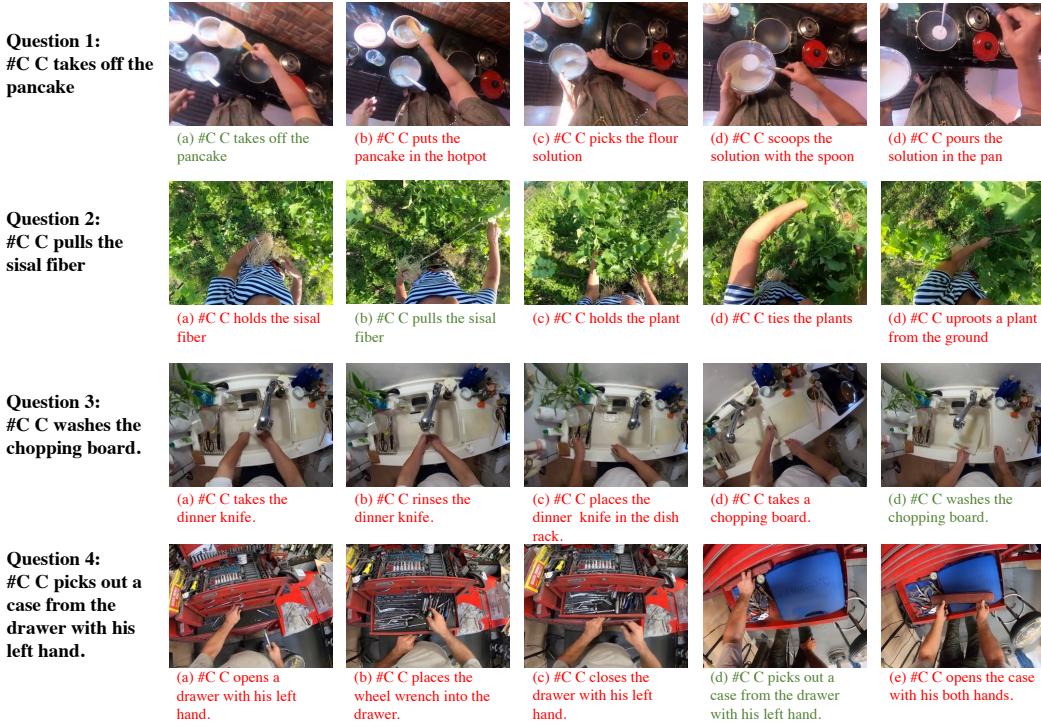


(g) #0 A man X moves hand from the table.

Figure 8: Visualization of EgoClip clip-text pairs. We sample five frames uniformly for each clip and take its narration as its caption.



(a) Inter-video setting



(b) Intra-video setting

Figure 9: 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 narrations is highlighted in green and the wrong in red.

B Construction details and statistics of EgoMCQ benchmark

B.1 Data deduplication

To ensure repetitions do not appear in five options, we devise a deduplication strategy. Initially, we use Sentence-BERT to extract sentence-level embeddings of narrations and set a manual threshold to remove repetitions. But in this way, it is hard to control the fine-grained diversity between narrations, e.g., two narrations “#C C closes the refrigerator with his left hand.” and “#C C opens the refrigerator with his left hand.” only differ in one word. These two sentences have a high score in sentence-level similarity, but are entirely different in semantic meanings. We hope to keep them and let the model distinguish them, especially in our intra-video setting.

Therefore, we propose to extract the first verb and the first noun of each narration and use them to define a tag for each narration. The narrations shared with the same verb and the noun will be assigned the same tag. We also consider the words synonyms (based on Ego4D taxonomy dictionary [16]). For instance, “#C C take the phone” and “#C C pick the cellphone” are semantically same in verb and noun thus will be assigned the same tag. Then the narrations shared with the same tag are treated as repetitions, we only keep one of them and sample a new one until the tags of the five options are different.

B.2 Data analysis

We finalize 39K questions covering 198K narrations with 468 hours of video, where the “inter-video” has 24K questions covering 290.3 hours of videos. And the “intra-video” has 15K questions and covers 178.3 hours of videos. The average duration among the five options is 34.2 seconds.

Geographic diversity. We present the geographic diversity of EgoMCQ in Fig 10, which covers 13 institutions and is align with the geographic diversity of EgoClip.

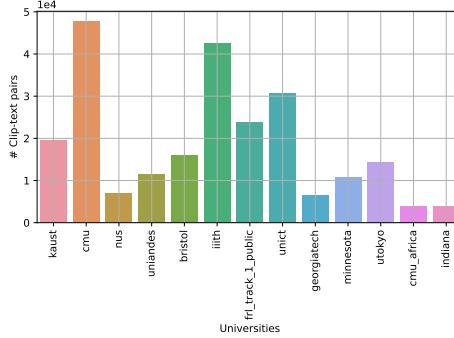
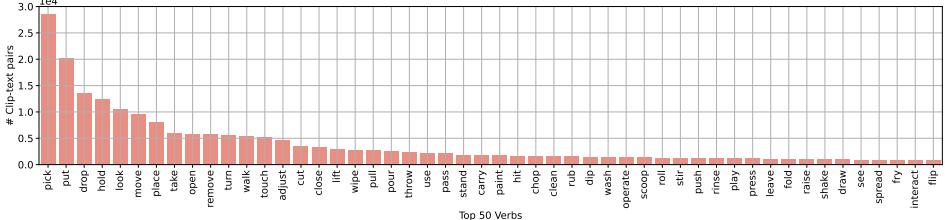


Figure 10: Institution distributions of EgoMCQ

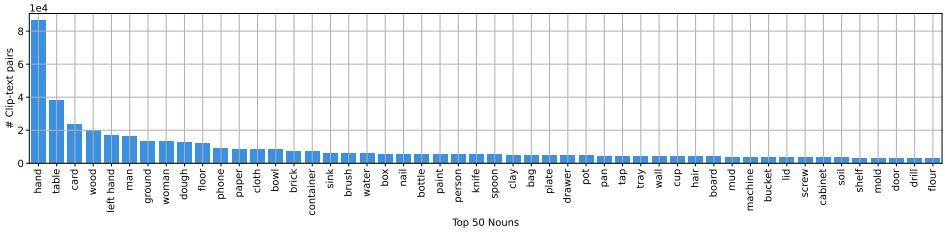
Scenario diversity. In Fig 5, we present the scenario distribution of EgoMCQ, which covers 74 scenario. The largest scenario “Cooking” includes 49K (15.3%) clips and the smallest scenario “Bus” contains 6 instances. EgoMCQ covers 71% of scenarios in EgoClip and has other 3 scenarios not appear in EgoClip. EgoMCQ is close to EgoClip both in terms of geography and scene diversity, making it a good development set for EgoClip pretraining.

Verbs and Nouns. EgoMCQ covers 198K narrations and each narration contains 3.15 nouns and 0.97 verbs in average. In Fig 11, we display the top 50 most frequently verb and nouns of EgoMCQ. The mostly common noun is “hand”, covering 86K (36.2%) instances and the mostly frequently verb is “pick”, which covers 28K (12.0%) instances.

Visualization. In Fig 9, we display examples of both the intra and inter settings of EgoMCQ respectively.



(a) Top 50 most frequently verbs distribution



(b) Top 50 most frequently nouns distribution

Figure 11: Verbs and nouns distributions of EgoMCQ’s narration.

C Technical details of our VLP model

In this section, we present more technical details of our VLP model, mainly architecture and pretraining objective.

C.1 Architecture: Frozen-in-time [11]

Video Encoder. The video encoder is built upon Timesformer [31], a convolution-free video backbone that divides space-time attention in an efficient manner. Take a RGB video clip $\mathcal{V}_i \in \mathbb{R}^{T \times 3 \times H \times W}$ with T frames and resolution $H \times W$ as input, the clip is first divided into $M \times N$ patches $\mathbf{p} \in \mathbb{R}^{M \times N \times 3 \times P \times P}$ with size of $P \times P$, where $N = HW/P^2$. Next, patches \mathbf{p} are linearly embed as a token sequence $\mathbf{z} \in \mathbb{R}^{MN \times D}$ with D dimension. Then, the learned temporal embeddings $E^s \in \mathbb{R}^{N \times D}$ and spatial positional embeddings $E^t \in \mathbb{R}^{N \times D}$ are added to each input token. Besides, a learnable CLS token is concatenated at the beginning of the token sequence. Finally, these token sequences are fed into Timesformer and output the CLS token of the last block, which is further projected to a d dimension embedding by a linear layer to form the final clip representation $\mathbf{v}_i \in \mathbb{R}^d$.

Text Encoder. The text encoder is built upon DistillBERT [32], which has 40% less parameters than BERT while also preserves over 95% performance, thus is efficient. Taking a sentence \mathcal{T}_i as input, first tokenize it as a sequence of tokens and feed it into DistillBERT. Similar to the video encoder, the CLS token of DistillBERT’s output is projected as $\mathbf{t}_i \in \mathbb{R}^d$ for the final text representation.

C.2 Pretraining objective: EgoNCE

To supplement the Eq. (2) and Eq. (3), we first formulate the complete form of InfoNCE:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{v2t}} + \mathcal{L}_{\text{t2v}} \\ &= \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\mathbf{v}_i^T \mathbf{t}_i / \tau)}{\sum_{j \in \mathcal{B}} \exp(\mathbf{v}_i^T \mathbf{t}_j / \tau)} + \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\mathbf{t}_i^T \mathbf{v}_i / \tau)}{\sum_{j \in \mathcal{B}} \exp(\mathbf{t}_i^T \mathbf{v}_j / \tau)} \end{aligned} \quad (4)$$

and our EgoNCE extends the above as Eq.5 via two sampling strategies:

$$\begin{aligned}\mathcal{L}^{\text{ego}} &= \mathcal{L}_{\text{v2t}}^{\text{ego}} + \mathcal{L}_{\text{t2v}}^{\text{ego}} \\ &= \frac{1}{|\tilde{\mathcal{B}}|} \sum_{i \in \tilde{\mathcal{B}}} \log \frac{\sum_{k \in \mathcal{P}_i} \exp(\mathbf{v}_i^T \mathbf{t}_k / \tau)}{\sum_{j \in \mathcal{B}} (\exp(\mathbf{v}_i^T \mathbf{t}_j / \tau) + \exp(\mathbf{v}_i^T \mathbf{t}_{j'} / \tau))} \\ &\quad + \frac{1}{|\tilde{\mathcal{B}}|} \sum_{i \in \tilde{\mathcal{B}}} \log \frac{\sum_{k \in \mathcal{P}_i} \exp(\mathbf{t}_i^T \mathbf{v}_k / \tau)}{\sum_{j \in \mathcal{B}} (\exp(\mathbf{t}_i^T \mathbf{v}_j / \tau) + \exp(\mathbf{t}_i^T \mathbf{v}_{j'} / \tau))}.\end{aligned}\tag{5}$$

For positive sampling (the numerator term), we pre-extract the nouns and verbs for each narration \mathcal{T}_i before pretraining and define two word vectors $\mathbf{w}_i^n \in \{0, 1\}^{K_1}$ and $\mathbf{w}_i^v \in \{0, 1\}^{K_2}$ to encode the appearing nouns and verbs in sentence, where K_1 and K_2 denote the number of nouns and verbs in EgoClip (Refer to Sec A narration analysis). During pretraining, for another instance j within batch, we calculate the $s_{ij} = (\mathbf{w}_i^n)^T \mathbf{w}_j^n \cdot (\mathbf{w}_i^v)^T \mathbf{w}_j^v$, if $s_{ij} > 0$, we regard instance j is one of the positive sample $j \in \mathcal{P}_i$ of instance i . Notably, the positive sampling space \mathcal{P} would cover $\tilde{\mathcal{B}}$ when working with the negative sampling strategy.

For negative sampling (the denominator term), each time we sample an instance i , we sample an instance $i' \in \mathcal{V}_i$ in the same video and close in time (less than 1 min) to generate the negative sample $i' \in \mathcal{N}(i)$ of instance i . Notably, in this way, the actual instance within the batch $|\tilde{\mathcal{B}}| = 2N$ will be double the batch size $|\mathcal{B}| = N$. In practice, we have to halve the batch size due to GPU memory limitations. Under halving the batch size, random sampling doesn't help in our method, which can be concluded by comparing baseline InfoNCE and variants (d) in Tab. 3 of the main body, where the batch size of the latter is half of the former. Despite this, our proposed sampling strategy (f) can successfully improve the pretraining effect beyond baseline.

D Additional experimental details and results

D.1 Implementation details

Following the settings of official Frozen [11], the video encoder is initialized with ViT [48] weights trained on ImageNet-21K with sequence dimension $D = 768$. The text encoder is based on huggingface's `distilbert-base-uncased`. The dimension of common feature space is set as 256, and the temperature parameter is set to 0.05. During pretraining, each video is resized to 224×224 as input with sample frames number 4 and batch size 512. We use the Adam optimizer with a learning rate of 3×10^{-5} with a total epoch of 10. When transferring to downstream tasks, we select the checkpoints with the best score on EgoMCQ benchmark i.e. average accuracy of inter-video and intra-video settings by default.

D.2 Downstream settings

We present the setting details of the downstream tasks we evaluated. For a fair comparison, for VLPs variants pretrained on different datasets, we use the same settings on downstream tasks, such as the fine-tuning objective.

EPIC-KITCHENS-100 text-video retrieval. In this task, after we finalize video-text pretraining, we continue to fine-tune the VLP model and keep most settings of pretraining (e.g., input resolution, learning rate). Notably, we set the training epoch as 100 and replace the training objective as Multi-instance Maxmargin loss in Eq.6, which is same as the baseline method JPSE [40]. The reason for this is that in this task a narration may be jointly associated with multiple clips, so multi-instance learning mechanism can better handle such a situation. And this dataset also provides the action label to calculate the correlation c_{ij} between two clip-text pairs (i, j) , which supports the implementation of Multi-instance Maxmargin loss.

$$\mathcal{L} = \sum_{(i,j,k) \in \Omega} \max(\gamma + \mathbf{v}_i^T \mathbf{t}_j - \mathbf{v}_i^T \mathbf{t}_k) + (\gamma + \mathbf{t}_i^T \mathbf{v}_j - \mathbf{t}_i^T \mathbf{v}_k), \tag{6}$$

where $\Omega = \{(i, j, k) | j \in i^+, k \in i^-\}$ is a triple, which indicates a positive instance j and a negative instance k for i . In our setting, we define the positive set as $i^+ = \{j | c_{ij} > 0.1\}$ and the negative as the remains sample within batch. The γ is a margin factor and we set it as 0.2.

Charades-Ego action recognition. In this task, the textual categories are short phrases like ‘‘Holding some clothes’’. Thus, we regard this task as a kind of video-text retrieval by leveraging the text representation and using the InfoNCE as fine-tuning objective. We set the epoch number as 10 and keep other parameters unchanged.

Natural Language Query This task is a kind of video-text localization and is hard to perform end-to-end training (since a clip might long to 1200 seconds). The baseline method [44] takes 2304 dim SlowFast features (1.87 fps, with Kinetics 400 pretrained) and 768 dim BERT features as input. Therefore, we propose to replace the baseline input features as features of pretrained VLP video and text encoders to evaluate the pretraining effectiveness. We extract the features with the same fps 1.87 and sampling frame number 4. In fine-tuning stage, we keep the default setting of [44].

Moment Query This task is a video-only task: temporal action localization. Similar to Natural Language Query task, we replace the input Slowfast features of baseline VSGN [45] with VLP video features for evaluation. The extraction details are the same as Natural Language Query.

Object State Change Classification This is an action classification task, we sample each clip with 16 frames as input and use the cross-entropy as fine-tuning objective. The epoch is set as 10.

D.3 VLP Evaluation on EgoMCQ

In Tab. 9, we display EgoMCQ evaluation result of Frozen pretrained on different video-text datasets.

VL Pretraining	Accuracy (%)	
	Intra-video	Inter-video
Random	20.0	20.0
EPIC-KITCHENS-100	28.1	22.7
HowTo100M	31.5	21.6
CC3M+WebVid2M	62.5	27.4
EgoClip	89.4	51.5
EgoClip w/ EgoNCE	90.6	57.2

Table 9: Results of VLPs pretrained on different datasets in EgoMCQ

As shown, pretraining with EPIC-KITCHENS-100 dataset (1st-person view, 67.2K pairs) reach comparable performance with HowTo100M pretraining (3rd-person view, 136M noisy pairs), which demonstrates the major domain gaps. Besides, Frozen with CC3M+WebVid-2M pretraining reach significant improvement on the intra-video setting, but minor in inter-video. We speculate this due to CC3M+WebVid-2M dataset covering a wide range of appearance information but still less exploration in the fine-grained action e.g. human-object interaction.

D.4 Training Curves of EPIC-KITCHENS-100 text-video retrieval

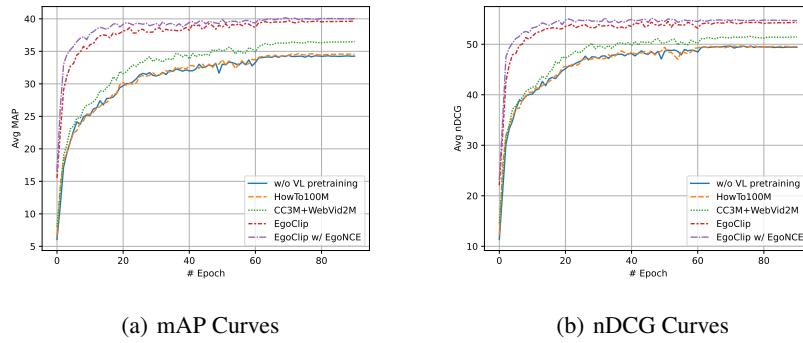


Figure 12: Training Curves of EPIC-KITCHENS-100 text-video retrieval

In Fig 12, we display training curves of EPIC-KITCHENS-100 text-video retrieval under different video-text pretraining, which also includes a baseline without video-text pre-training. We can find that: Variants with video-text pretraining have a faster rise in performance. Except for HowTo100M, which is similar to variant without video-text pretraining. Especially with EgoClip for egocentric pretraining, the VLP model achieves nearly convergent performance with only a small number of epochs (less than 20). With EgoNCE as pretraining objective, this positive effect is further enhanced.

D.5 Results on test set of Moment Query

We further display the test set results of Moment Query in Tab. 10, pretraining with EgoClip and EgoNCE reach the best performance, 3.78% on R@1 and 4.65% on Avg mAP over the baseline.

Methods	Video-text Pre-extracted Features		IoU=0.5 R@1	mAP(%) Avg
	Vis-text Enc	Vis-text PT		
VSGN	SlowFast	-	24.25	5.68
VSGN	Frozen	HowTo100M	18.06	5.28
VSGN	Frozen	CC3M+WebVid2M	19.74	5.95
VSGN	Frozen	EgoClip	<u>27.98</u>	<u>9.78</u>
VSGN	Frozen+EgoNCE	EgoClip	28.03	10.33

Table 10: Recall and mAP metrics on the MQ task’s test set.

D.6 Results on test set of Natural Language Query

In Tab. 11, we found the similar conclusions in test set of Natural Language Query, pretraining with EgoClip and EgoNCE reach the optimum performance.

Methods	Video-text Pre-extracted Features		IoU=0.3		IoU=0.5	
	Vis-text Enc	Vis-text PT	R@1	R@5	R@1	R@5
VSLNet	SlowFast+BERT	-	5.47	11.21	2.80	6.57
VSLNet	Frozen	HowTo100M	3.77	6.87	1.62	3.45
VSLNet	Frozen	CC3M+WebVid2M	4.87	8.67	2.50	4.97
VSLNet	Frozen	EgoClip	<u>10.34</u>	<u>15.81</u>	<u>6.24</u>	<u>10.39</u>
VSLNet	Frozen+EgoNCE	EgoClip	10.46	16.76	6.24	11.29

Table 11: Recall for several IoU on the NLQ task’s test set.

D.7 Visualization

To intuitively understand the effect of egocentric pre-training, in Fig 13, we compare the EPIC-KITCHENS-100 text-video retrieval results between our pre-training (EgoClip w/ EgoNCE) and CC3M+WebVid-2M pre-training, both fine-tuning with 16 frames. The numbers after each narration represent the correlation scores between the query and the retrieval result, with 1 being the best.



Figure 13: Visualization of EPIC-KITCHENS-100 text-video retrieval. Given the same text query, we compare the **Top-5 results** of 1st-person pretraining and 3rd-person pretraining.