Learning Temporal Video-Language Grounding for Egocentric Videos

Presenter: Priyanka Mandikal, Alex Chandler, and Liyan Tang

04-24-2023

Research Problem

Grounding video-language embeddings to learn the temporal structure of long-horizon egocentric videos

Motivation

- Egocentric video understanding → wide applications
- Downstream computer vision and language applications
  - Augmented reality
  - Robot learning
- Example:
  - "Where did I leave my keys?"
  - "Did I turn off the stove before leaving the house?"
- Our work → mapping natural language queries to segments in egocentric videos

Example downstream task of EgoVQA from Google

Challenges

- Egocentric videos → inherently complex
  - High degree of diversity in actions, environments, and camera movements
- Long-horizon videos → large contexts
  - Challenging to identify relevant information from lengthy multi-segment videos
- Natural language queries can be ambiguous
  - Require understanding of both context and temporal relationships
- Large-scale training → computationally challenging

Prior Work: Video Language Pretraining

- Goal: Learn useful representation of videos and their associated textual information
- Downstream tasks: video-text retrieval, video question answering and video captioning.

Prior Work: Video Language Pretraining

- Goal: Learn useful representation of videos and their associated textual information
- Downstream tasks: video-text retrieval, video question answering and video captioning.
- Popular video-language pre-training techniques:
  - Masked Language Modeling (MLM)
    - Masking textual information, video features or both simultaneously
  - Frame Order Modeling (FOM)
    - Temporal reasoning → predict the original order of a sequence of shuffled frames
  - Video Language Matching (VLM)
    - Contrastive learning → on positive and negative samples

Prior Work: Video Language Pretraining

- Goal: Learn useful representation of videos and their associated textual information
- Downstream tasks: video-text retrieval, video question answering and video captioning.
- Popular video-language pre-training techniques:
  - Masked Language Modeling (MLM)
    - Masking textual information, video features or both simultaneously
  - Frame Order Modeling (FOM)
    - Temporal reasoning → predict the original order of a sequence of shuffled frames
  - Video Language Matching (VLM)
    - Contrastive learning → on positive and negative samples

Our approach → combine best of both FOM and VLM techniques

Prior Work: Video-Text Retrieval

- Goal: Locate and identify specific video segments or moments that match the given natural language query or description.

Prior Work: Video-Text Retrieval

- Goal: Locate and identify specific video segments or moments that match the given natural language query or description.

- Step 1: Extract spatio-temporal visual and textual features from the input video and query

- Step 2: Establishing a correspondence between them to identify relevant video segments
  - Global matching
    - Capture the overall semantic similarity between the video and textual representation
  - Local matching
    - Focus on the similarity between fine-grained local features from video and text

Prior Work: Video-Text Retrieval

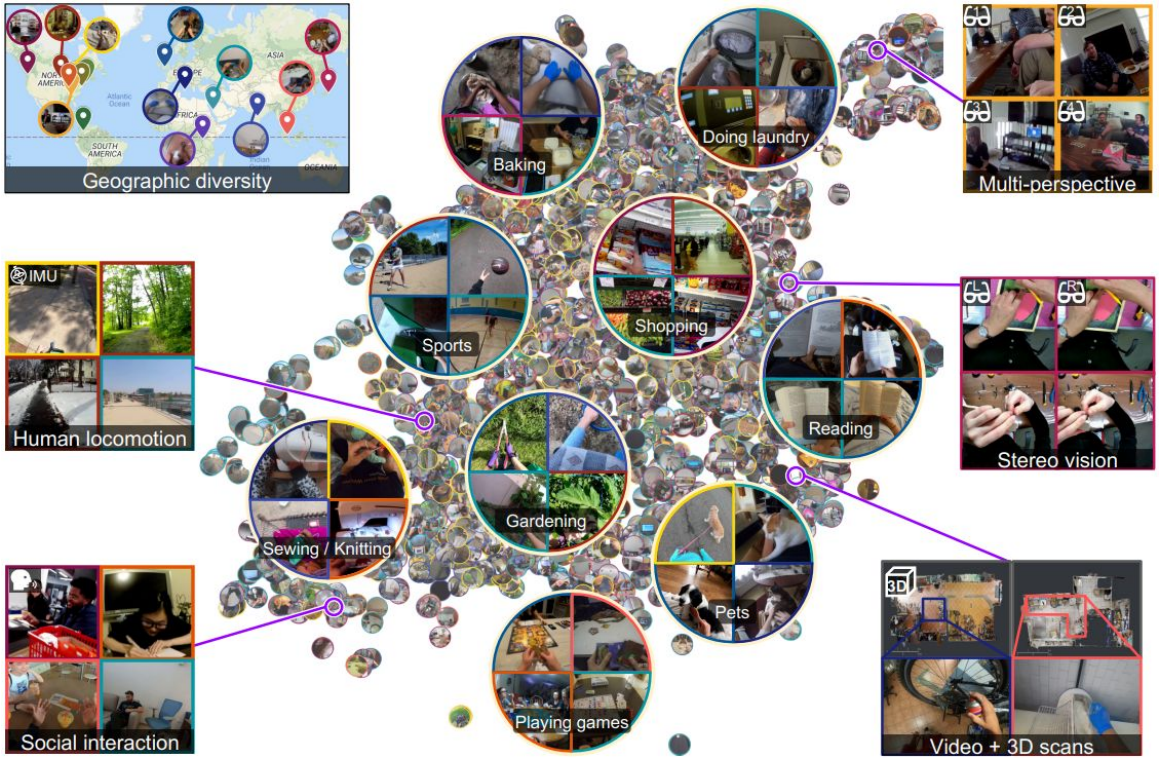
- Goal: Locate and identify specific video segments or moments that match the given natural language query or description.

- Step 1: Extract spatio-temporal visual and textual features from the input video and query

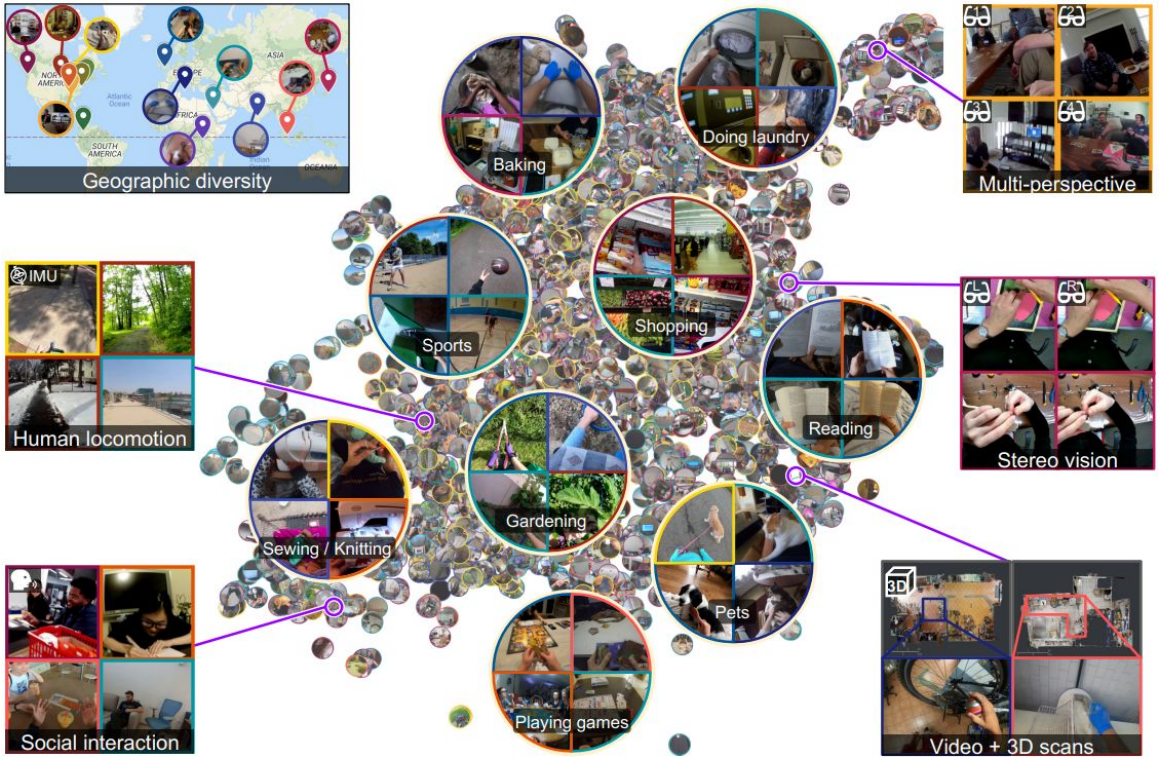
- Step 2: Establishing a correspondence between them to identify relevant video segments
  - Global matching
    - Capture the overall semantic similarity between the video and textual representation
  - Local matching
    - Focus on the similarity between fine-grained local features from video and text

Our work → global matching: CL for strong global repr of visual + textual features

Dataset: Ego4D



Dataset: Ego4D



Diverse

Multimodal

Large-Scale

User Bias

Dataset: EgoCLIP

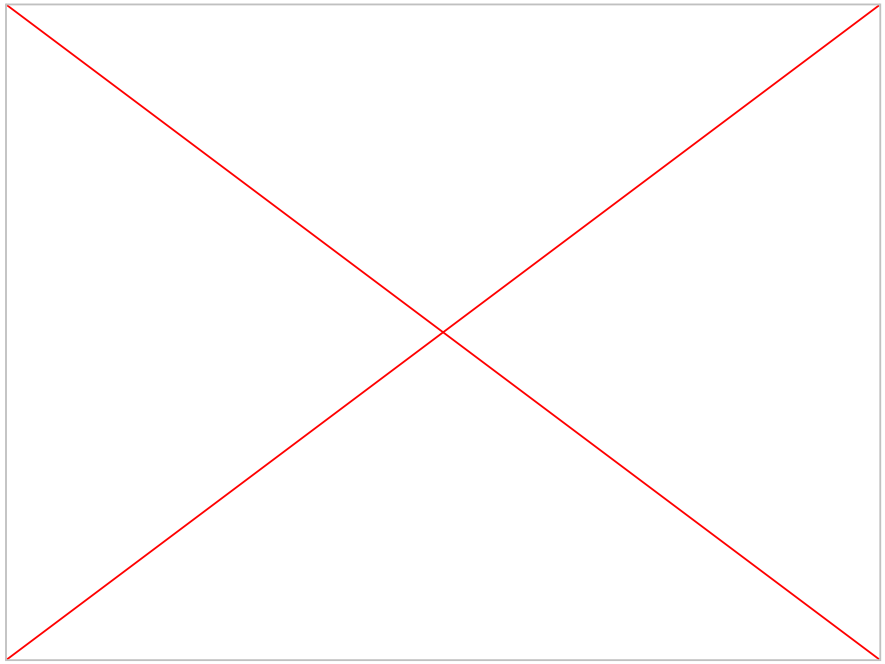
- Train and evaluate on **EgoCLIP** dataset
  - Subset of Ego4D used for pre-training video-language models
- 3.8 million video-text pairs → cover a wide range of daily human activities from first-person POV
- Dense timestamp-level narrations from multiple separate narrators

Comparison of EgoCLIP Dataset

Dataset	Ego?	Domain	Dur (hrs)	# Clips	# Texts
MSR-VTT [1]	✗	diverse	40	10K	200K
YouCook2 [16]	✗	cooking	176	14K	14K
ActivityNet Captions [7]	✗	action	849	100K	100K
WebVid-2M [3]	✗	diverse	13K	2.5M	2.5M
HowTo100M [10]	✗	instructional	134K	136M	136M
Charades-Ego [17]	✓	home	34	30K	30K
UT-Ego [18]	✓	diverse	37	11K	11K
Disneyworld [19]	✓	disneyland	42	15K	15K
EPIC-KITCHENS-100 [20]	✓	kitchen	100	90K	90K
EgoClip	✓	**diverse**	**2.9K**	**3.8M**	**3.8M**

Source: EgoCLIP Lin 2022

EgoCLIP Video Example

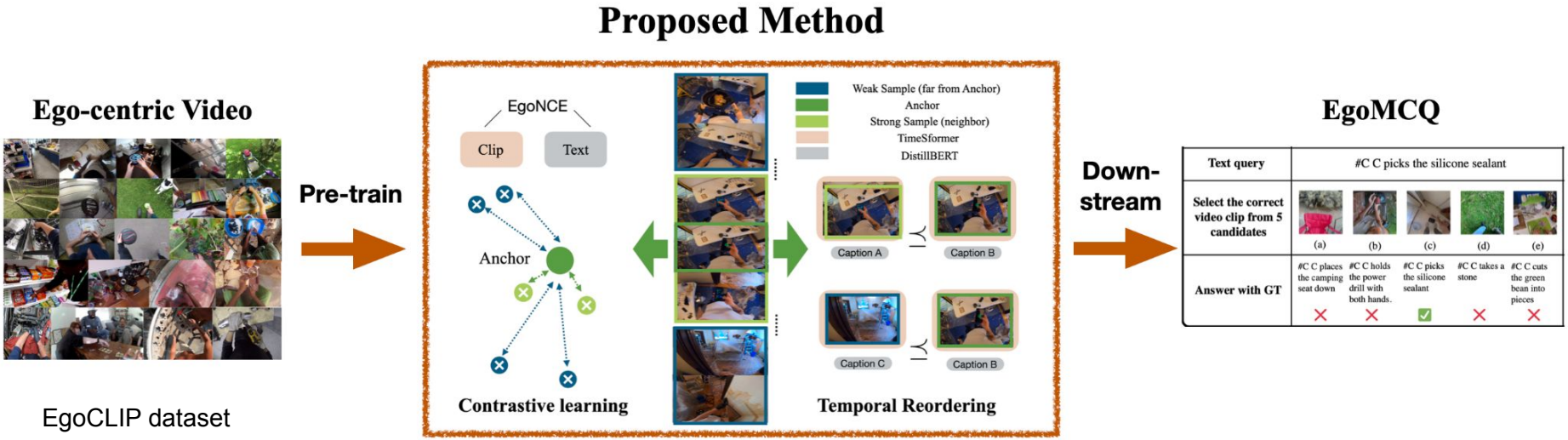


Video Clip Narration Texts

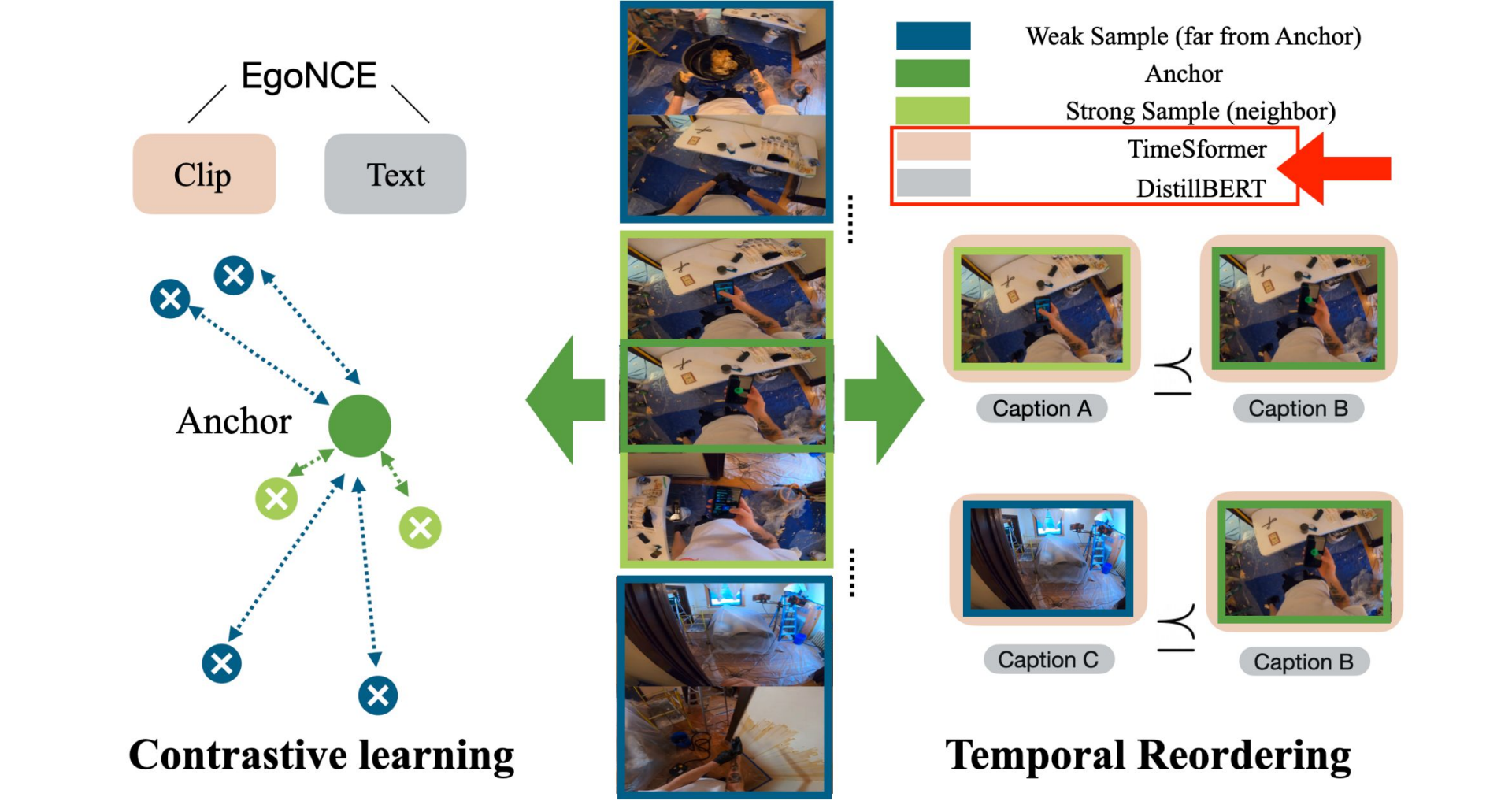
- Each video (avg 20 minutes long) consists of multiple clips (avg 1.0 sec long)
- Clip Start, Narration 1 Pairings

```
[[0.0, 'C stares at the room'],
[0.5, 'C adjusts the camera on the head'],
[4.1, 'A man X walks around the room'],
[21.0, 'C picks up the glove'],
[22.4, 'C puts on the glove'],
[26.6, 'C looks around the room'],
[27.8, 'A man X plays with the tape'],
[29.9, 'C puts on the glove'],
[33.9, 'C picks up the glove'],
[36.7, 'C puts on the glove'],
[55.0, 'C walks around the room'],
[60.7, 'C looks around the room'],
[63.2, 'C picks the wallpaper remover'], …]
...
```

Overall Pipeline

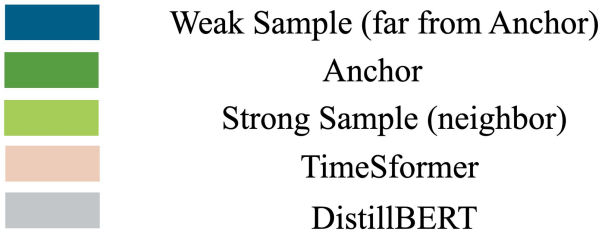


EgoCLIP dataset

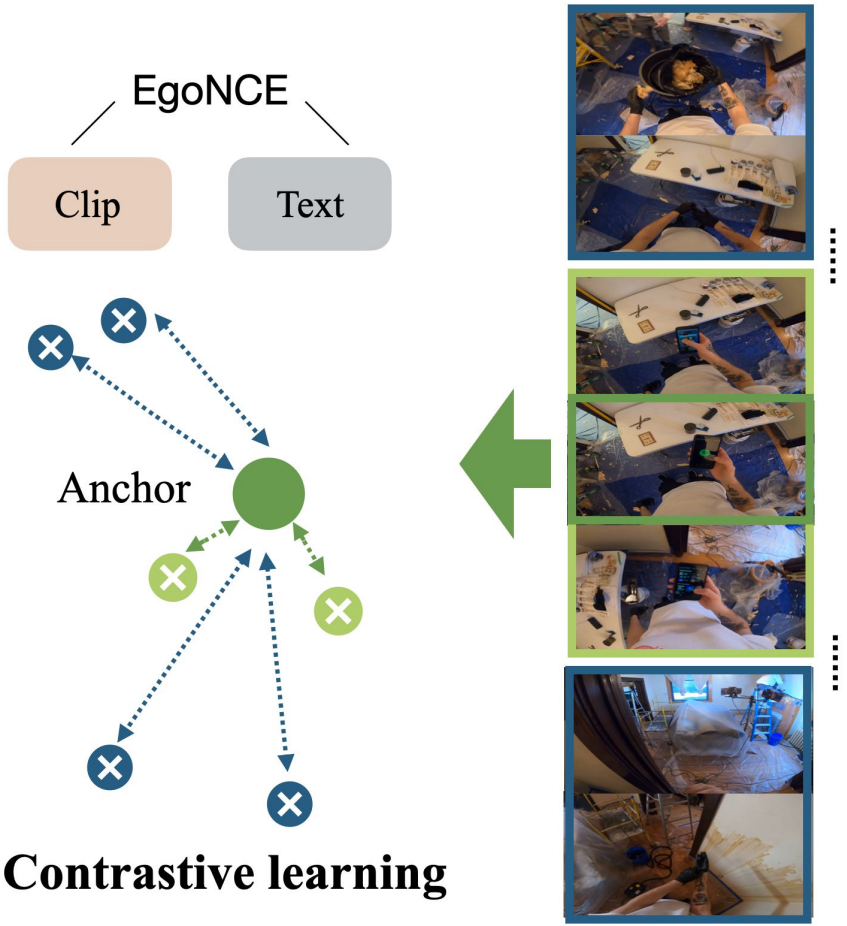

EgoNCE
Clip
Text
Anchor
Contrastive learning
Weak Sample (far from Anchor)
Anchor
Strong Sample (neighbor)
TimeSformer
DistillBERT
Caption A
Caption B
Caption C
Caption B
Temporal Reordering

Proposed Methodology

(Negative Example) Sampling for Contrastive Loss



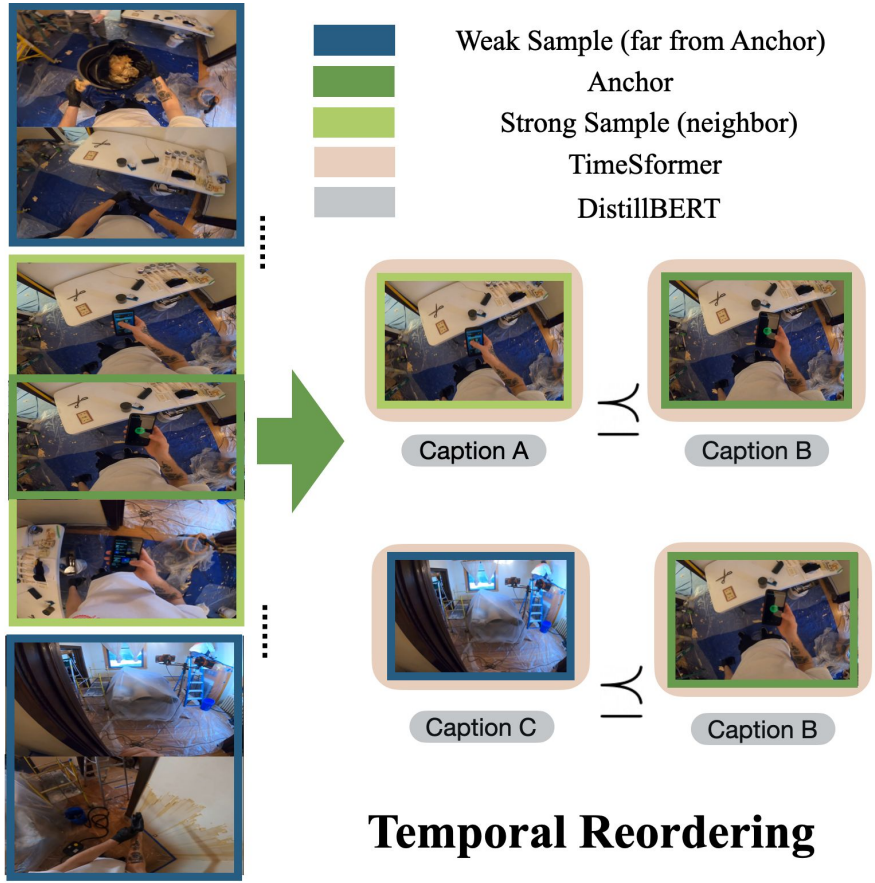
- Weak Contrastive Loss Objective
  - Samples far from the anchor
- Strong Contrastive Loss Objective
  - Samples from the neighbor of the anchor



Proposed Methodology

Temporal Reordering Loss Objective for Video Clips

- Weak reordering
  - Samples to reorder are far away
- **Strong reordering**
  - Samples to reorder are from the neighbor of the anchor

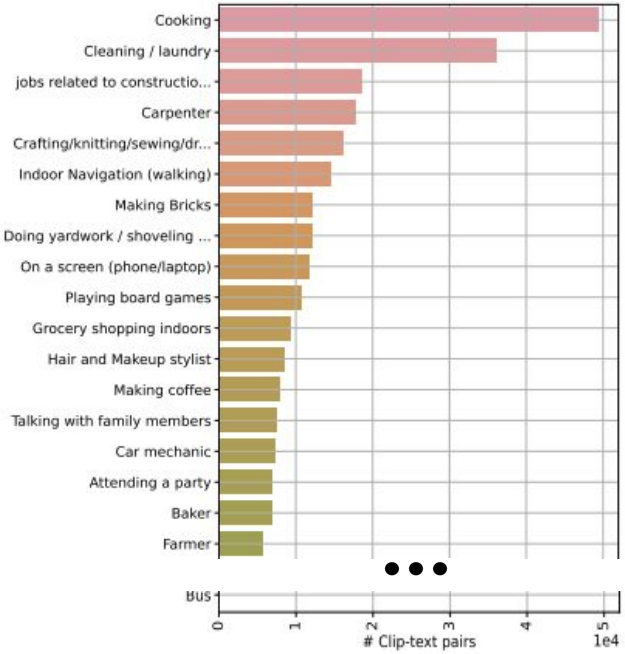


Evaluation Metrics

We evaluate on the **EgoMCQ** benchmark

- Measures video-text alignment in egocentric videos with questions from the Ego4D dataset.
- Consist of 39k questions, 198k narrations, and 468 hours of video.
- Evaluate on **accuracy**
  - Probability of selecting correct answer from 5 multiple choices) under both inter-video (answer choices from different video) and intra-video (answer choices from same video).

Figure 5: Scenario distribution of EgoMCQ



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	 (a)	 (b)	 (c)	 (d)	 (e)	 (a)	 (b)	 (c)	 (d)	 (e)
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 ✓

Given a text query, pick the correct image that corresponds to our query out of 5 multiple choice images.

Experiments

- Baseline: Train with using EgoVLP Sampling Technique
  - Scene-aware Negative Sampling applied in EgoNCE loss
- Train with Strong and Weak Sampling
  - 50% Strong 50% Weak
- Train with Strong Sampling and Temporal Reordering

Training Details

- Train on 10% of dataset
- Train each experiment with 5 epochs

Results

Method	Training Data	Epochs	Sampling Strategy	Temporal Reordering	EgoMCQ Acc (%)	
					Intra-Video	Inter-Video
EgoVLP	100 %	10	Random	No	89.4	51.5
EgoVLP	10 %	5	Random	No	82.9	46.2
Ours	10 %	5	Strong and Weak	No	84.0	**47.3**
Ours	10 %	5	Strong	Yes	**84.3**	**47.3**

Model Architecture Assessment

- Strengths
  - Two-fold approach to training that leverages both hard negatives of neighbouring video clips and a temporal reordering objective in the contrastive loss to improve the model's ability to learn temporal ordering of video-language embeddings.
  - TimeSFormer architecture allows for benchmark comparison on downstream tasks in EgoVLP
- Limitations
  - Our architecture, even after temporal reordering, may not capture long-term temporal dependencies to understand the egocentric activities in the videos.
  - Could not train on full EgoVLP dataset due to computational demands from TimeSFormer

Summary

- Sampling strategies for contrastive loss function are important
  - The use of hard negatives from nearby video segments in the same video serves as a form of curriculum learning, forcing the model to differentiate between more subtle differences in the visual field and improving its ability to generalize to new data.
- Empirical signs of improvement by introducing temporal reordering objective in the contrastive loss
- The model's increased ability to capture complex temporal dependencies between actions could improve its performance in downstream tasks such as natural language querying, which relies on accurate temporal understanding of the underlying video content.

Future Work

- Planned Future Work
  - Downstream Evaluation
    - EgoNLQ
    - Use fine-tuned features on NLQ task
    - EgoVQA
    - etc.
  - Improved Temporal Reordering
    - Need for hyperparameter tuning
  - Explore newer Video Transformers
    - ex: VideoMAE, VideoMAEV2

Extended Readings

Helpful Readings:

- **Papers:**
  - TimeSFormer
  - EgoVLP
  - MERLOT
  - LaViLa
- **Blogs**:
  - Understanding TimeSFormer
  - Noise Contrastive Estimation

Other Readings:

- Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning
- Object-aware Video-language Pre-training for Retrieval
- VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

Thank You