Weakly Supervised Video Representation Learning with Unaligned Text for Sequential Videos

Sixun Dong[1*], Huazhang Hu[1*], Dongze Lian[1,2], Weixin Luo[3], Yicheng Qian[1], Shenghua Gao[1,4,5†]
[1]ShanghaiTech University [2]National University of Singapore [3]Meituan
[4]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[5]Shanghai Engineering Research Center of Energy Efficient and Custom AI IC
{dongsx, huhzh, liandz, luowx, qianyc, gaoshh}@shanghaitech.edu.cn

Abstract

Sequential video understanding, as an emerging video understanding task, has driven lots of researchers' attention because of its goal-oriented nature. This paper studies weakly supervised sequential video understanding where the accurate time-stamp level text-video alignment is not provided. We solve this task by borrowing ideas from CLIP. Specifically, we use a transformer to aggregate frame-level features for video representation and use a pre-trained text encoder to encode the texts corresponding to each action and the whole video, respectively. To model the correspondence between text and video, we propose a multiple granularity loss, where the video-paragraph contrastive loss enforces matching between the whole video and the complete script, and a fine-grained frame-sentence contrastive loss enforces the matching between each action and its description. As the frame-sentence correspondence is not available, we propose to use the fact that video actions happen sequentially in the temporal domain to generate pseudo frame-sentence correspondence and supervise the network training with the pseudo labels. Extensive experiments on video sequence verification and text-to-video matching show that our method outperforms baselines by a large margin, which validates the effectiveness of our proposed approach. Code is available at https://github.com/svip-lab/WeakSVR.

1. Introduction

A strong artificial intelligence (AI) system is expected to be able to learn knowledge from the open world in an embodied manner such that amounts of goal-oriented tasks are designed for reinforcement learning in the environment. In the area of video understanding, a great deal of pioneering work in video classification [56], action localization [54], and action segmentation [26] has been explored, laying the foundation for video understanding. Beyond these typical video understanding tasks, sequential videos (such as Fig. 1) that usually describe how to perform a task in a certain sequence of procedures can be regarded as a goal-oriented task. Solving this task is extremely promising for guiding intelligence to learn a task like humans. It makes performing sequential video representations a potentially critical part of the road to strong AI.

Some efforts have been made for video representation learning for sequential videos. *e.g.*, [1, 18] learns a video representation in an instructive video. However, these methods rely heavily on the annotations of temporal boundaries, i.e., the timestamps of sequential actions, which are usually difficult to be obtained due to the time-consuming human labeling in practice. A common but often overlooked scenario is that sequential videos usually occur accompanied with audio or text narrations, which show consistent steps with explanations. The rich text information describes the corresponding procedure in detail as shown in Fig. 1, but they are usually not aligned with videos. Therefore, a question arises, i.e., whether it is possible to directly learn the video representation with unaligned text and video in a weakly supervised manner.

With the popularity of visual-language tasks, multimodal learning has attracted growing attention and has been explored in a variety of areas, e.g., image classification [5, 50], object detection [42, 63], and video understanding [61]. One of the most representative works is CLIP [43]. It has shown the potential of learning a powerful semantic representation from natural language supervision with a contrastive learning loss and the strong zero-shot generalization on the downstream tasks, such as text-video retrieval [48, 58], action segmentation [65], multiple-choice videoQA [17, 58] and action step localization [4]. VideoCLIP [59] presents a contrastive learning approach to pre-

[*] Equal Contribution.
[†] Corresponding Author.

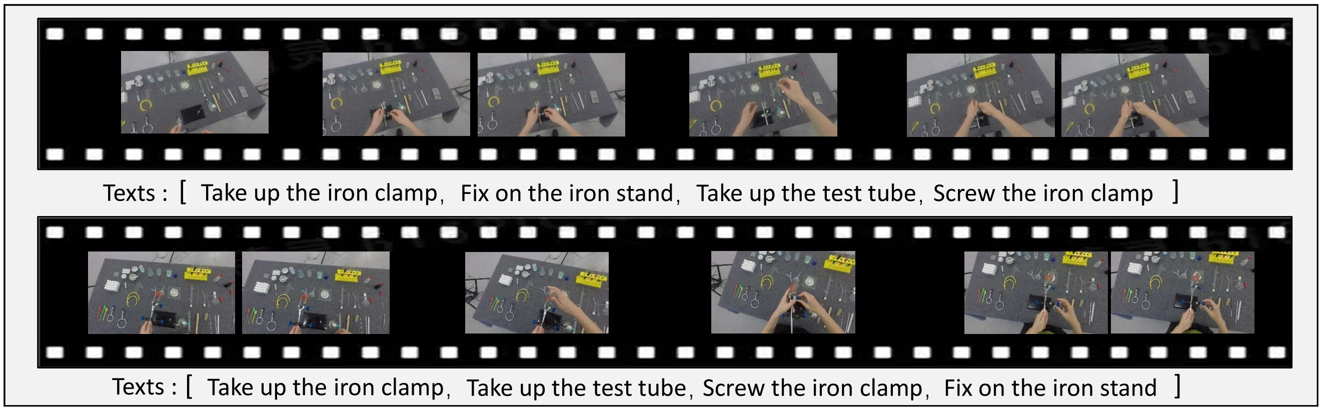

Texts : [Take up the iron clamp, Fix on the iron stand, Take up the test tube, Screw the iron clamp]
Texts : [Take up the iron clamp, Take up the test tube, Screw the iron clamp, Fix on the iron stand]

Figure 1. **Sequential Video.** The samples come from CSV dataset. They describe two types of step schedule to accomplish the task of ”fix the test tube on the iron stand with iron clamp”. The upper process the step ”fix on the iron stand ” before the steps ”take up the test tube” and ”screw the iron clamp”. Diversely, the lower make the steps of ”take up the test tube” and ”screw the iron clamp” before the step ”fix on the iron stand ”. It can be seen that the order, time span and temporal location of sub-actions to accomplish the task are apparently different.

train a unified model with video-text pairs, and [1] proposes a unified fully and timestamp-supervised framework for multi-model action segmentation. This provides us with an alternative for weakly supervised video representation learning. However, all these previous works are equipped with aligned texts and video frames [1], which is not existent in our weakly supervised setting. Thus, it is intractable to directly adapt the existing multi-modal video representation models to our task.

To overcome the unalignment issue between text and video and learn a satisfactory video representation, we propose a weakly supervised video representation learning pipeline and introduce a multiple granularity contrastive loss to constrain the model, which takes full account of the pseudo temporal alignment between frames and sentences. To be specific, we first extract video and text features from a CLIP-based vision-language model, and a global contrastive loss is designed to constrain the complete video-paragraph alignment. It constrains that a video will be closer to the sequence of the texts describing it while far away from the rest of the texts, and vice versa. Secondly, we introduce a fine-grained contrastive learning loss, which encourages the frame sequences of representations to be more similar to the neighbor sentence representations than the remote sentences in the same paragraph. The intuition behind this constraint comes from a basic idea: *if the s_j is the corresponding sentence for frame h_i, the corresponding sentence for frame h_{i+1} is never before the s_j in sequence*. Specifically, we take the probabilistic sample from the sentence-frame similarity metric. And we propose to apply the differentiable Gumbel-Softmax [23] tricks to generate predictions and propose three kinds of methods to generate the pseudo-labels that are based on the temporal relation of sentences in the temporal domain: 1) maximum-index sorting; 2) Viterbi algorithm [16]; 3) splitting. Finally, we calculate the Info-NCE contrastive loss based on the pseudo labels in order to guide the network to focus on the fine-grained action matching in sequential videos.

To evaluate the effectiveness of our weakly supervised video representation method, we conduct extensive experiments on two downstream tasks: video verification in procedures and text-to-video matching. The results of experiments show that our approach outperforms other baselines by a significant margin and also demonstrates the great generalization of our model.

We summarize our contributions in three folds:

- We propose a novel weakly supervised video representation learning pipeline with unaligned text for sequential videos, which can learn powerful and semantic video-text representations.
- We design multiple granularity contrastive learning loss, including coarse-grained loss and fine-grained loss. Notably, we propose a novel method to implement the temporal alignment between frames and sentences.
- Our model also shows strong generalization ability to downstream tasks, such as video sequence verification for procedures in videos and text-to-video matching.

2. Related Works

Sequential Video. The same task described in videos may consist of several sequential sub-actions in different orders for a sequential video. Sequential videos are generally accompanied by explanations such as audio or caption. Various kinds of studies related to sequential videos are now in the ascendant. For example, COIN [49], Diving [29], CSV [41], EPIC-KITCHENS [9], IKEA-ASM [2] and Assembly101 [44] provide videos composed by multiple sequential actions and the corresponding step annotations. [44] proposes a large-scale multi-view video dataset for understanding procedural activities, which is beneficial for the whole community. [41] defines the pioneering sequence verification task and designs a method based on the alignment of video pairs. However, the method is seriously dependent on video pairs of the same tasks. [31] learns to recognize procedural activities in sequential videos with distant supervision [38, 62]. [34] propose an action segmentation method using the set-supervised method for sequential videos. [25] employs temporal optimal transport to generate pseudo labels to complete joint representation learning and online clustering for sequential video alignment. D^{3}TW [6] aligns clips and transcripts with differentiable continuous relaxation.

Vision-text Multi-modality Learning. Vision-text multi-modality [4, 18, 37, 43, 46, 47, 53, 57, 65] has attracted increasing attention in computer vision communities over the recent year. One of the most representative works is CLIP [43], which is able to learn a powerful visual representation from natural language supervision with contrastive learning loss. Due to the strong zero-shot generalization ability of the method, a large number of follow-up works have been proposed [4, 18, 28, 39, 48, 57, 59]. VideoCLIP [59] presents a contrastive approach to pre-train a unified model with video-text pairs. X-CLIP [39] effectively expands the pre-trained language-image model to video domains based on a cross-frame attention mechanism. However, these methods heavily rely on strong data augmentation and a large batch size. For downstream tasks, LocVTP [4] shows its transfer potentials on localization-based and retrieval-based tasks. CLIP4Clip [35] uses the pretrained CLIP as our backbone to solve the video clip retrieval task from frame-level input. [17] bridges video-text retrieval with multiple-choice questions. LF-VILA [48] applies a multi-modality temporal contrastive loss to implement long-form video-language pre-training, which heavily relies on the timestamp annotations of clip-sentence pairs.

Video Representation Learning. Learning good video representations has been heavily investigated in the literature. 3D convolution neural networks (3D-CNNs) are originally considered to learn deep video representations [5, 14, 51]. However, 3D-CNNs are limited to capturing long-term dependencies on the temporal domain with their insufficient receptive field. Due to the ability to capture long-term dependency of the self-attention mechanism [52], vision transformer models [3, 7, 11, 13, 22, 32, 33] show competitive performances against 3D-CNNs in video representation learning. Following the ViT [11], many related works emerge. TimeSformer [3] designs different self-attention schemes in the temporal-spatial domain. Video Swin-Transformer [33] adopts the local attention in non-overlapping shifted windows to lead to a better speed-accuracy trade-off. Over recent years, weakly supervised or self-supervised learning [7, 45] is popular for learning better video representation. Following SimCLR [8], [7] introduces a self-supervised contrastive transformer framework to learn frame-wise action representations. [30] proposes a transformer-based cross-modal architecture for zero-shot action recognition. Previous works mainly focus on short-form simple video representation, whereas representation learning of sequential video is underexplored.

3. Method

In this section, we first present the overall architecture of the proposed framework in Sec. 3.1. Then we explain the vision representation module and language representation module in Sec. 3.2, followed by the designed multiple granularity contrastive learning module in Sec. 3.3 and Sec. 3.4.

3.1. Overview

Fig. 2 displays the overview of our framework. Our framework consists of three parts: a vision representation module, a language representation module, and a multiple granularity contrastive learning module. In the vision representation module, which shows in the right part of the figure, we sample frames from an untrimmed sequence video as input and extract visual features with the pre-trained vision encoder (unfrozen). After that, we concatenate the visual feature and pass them into the Transformer encoder. The Transformer encoder implements the cross-frame communication with self-attention and outputs the frame representations, following ViT [11]. Additionally, the results from Transformer encoder are then passed through the MLP module to integrate the frame representations and obtain the video representation. In the left language representation module, a collection of text descriptions of procedures and the description of the entire video pass into the pre-trained language encoder (frozen) separately, then we can obtain the sentence representations and a paragraph representation. More explanation about the aforementioned modules is in Sec. 3.2. Finally, we introduce multiple granularity contrastive loss to restrict learned representations in cross-model space.

3.2. Vision-Language Modules

As illustrated in Fig. 2, multi-level video representation and language representation are produced by the vision module and language module, respectively.

Vision module. Following [55], given an untrimmed sequence video, we uniformly split the raw video into N clips and randomly sample one frame per clip to form a sequence of N frames, $X = \{x_1, x_2, \dots, x_N\}$. Then we feed the frame sequence X into the pre-trained vision encoder E_v to

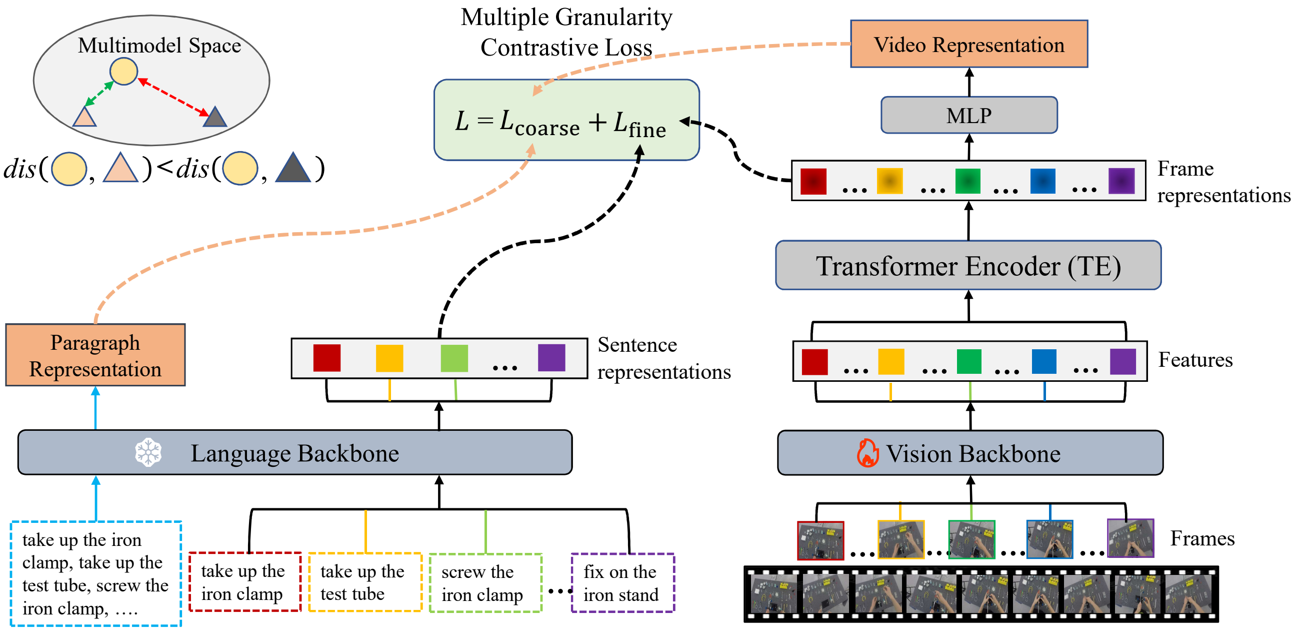


Figure 2. **Overview of our framework.** Our framework consists of three parts: vision representation module, language representation module, multiple granularity loss module. In the vision representation module, we feed the frames sampled from the untrimmed sequential video into the module, then obtain the frame representations and a video representation. In the language representation module, a collection of texts of procedures and the description of the entire video pass into the pre-trained language backbone separately, and we get the sentence representations and a paragraph representation. In addition, we introduce multiple granularity contrastive learning loss to restrict representations in cross-model space.

produce a sequence of feature maps $\{f_1, f_2, \dots, f_N\}$. This process can be denoted as $f_i = E_v(x_i), i \in [1, 2, \dots, N]$. After that, we prepend a learnable embedding x_{cls} to the sequence of features, called $[class]$ token [11]. Then as Eq. (1) shown, our method learns the frame representations by utilizing the transformer encoder (TE) to embed temporal and context information into frame representations $H = \{h_1, h_2, \dots, h_N\}$.

$$H = \text{TE}([x_{\text{cls}}, f_1, f_2, \dots, f_N] + e^{\text{pos}}) \tag{1}$$

where $[.,.]$ concatenates the features of frames and $[class]$ token. And e^{pos} represents the temporal position embedding of sequence.

At last, the MLP module, which consists of a full connection layer, takes all frame representations H as input and outputs a video representation v as follows:

$$v = \text{MLP}(H) \tag{2}$$

Language module. Specifically in our model, given a sequence of K text descriptions of procedures $T = \{t_1, t_2, \dots, t_K\}$, we first feed individual procedure texts into the frozen pre-trained language encoder E_l to produce sentence representations $S = \{s_1, s_2, \dots, s_K\}$. The process can be denoted as $s_i = E_l(t_i), i \in [1, 2, ..K]$.

In the meantime, we combine the sequence of text descriptions of procedures T into a single text description of the entire video. Then, the pre-trained language encoder E_l extract a paragraph-level language representation l as follows:

$$l = E_l([t_1, t_2, \dots, t_K]) \tag{3}$$

where $[.,.]$ represents simply the sequential concatenation of strings.

3.3. Coarse-grained Contrastive Loss

We first conduct contrastive learning at the video-paragraph level. Specifically, through the vision-language module that is explained in Sec. 3.2, we obtain a video representation V and paragraph representation L, where $V, L \in \mathbb{R}^{1 \times D}$. Then use one batch of data, $V = \{v_1, v_2, \dots, v_N\}$, $L = \{l_1, l_2, \dots, l_N\}$, to calculate the loss.

After that, we formulate the global video-paragraph alignment into the standard contrastive framework [43] based on InfoNCE loss [40] as follows:

$$L_{\text{InfoNCE}}(V, L) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\varphi(v_i, l_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\varphi(v_j, l_j)/\tau\right)} \tag{4}$$

$$\varphi(v_i, l_i) = \frac{v_i}{\|v_i\|} \cdot \frac{l_i^T}{\|l^T\|} \tag{5}$$

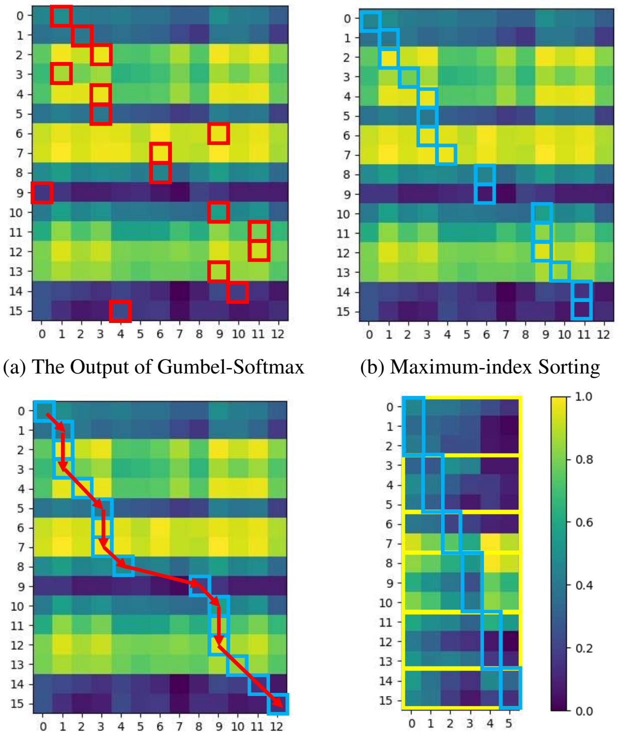
where τ is the temperature parameter optimized during training [43]. And $\varphi(.,.)$ represents the cosine similarity function, and N is the number of video-text pairs. The L_{InfoNCE} represents the InfoNCE loss.

Last, as shown Eq. (6), we calculate symmetrically video-text and text-video loss by Eq. (4) to obtain the coarse-grained contrastive loss L_{coarse}:

$$L_{\text{coarse}} = L_{\text{InfoNCE}}(V, L) + L_{\text{InfoNCE}}(L, V) \tag{6}$$

Showing in the upper left of Fig. 2, the coarse-grained global contrastive loss L_{coarse} restricts the representation in the cross-model latent space with video-paragraph level supervision.

3.4. Fine-grained Contrastive Loss



(a) The Output of Gumbel-Softmax

(b) Maximum-index Sorting

(c) Viterbi Algorithm

(d) Splitting

Figure 3. **Visualization** of fine-grained contrastive loss. The upper left figure shows the similarity matrix with Gumbel-Softmax. The other three figures show three kinds of pseudo-labels generation methods respectively: 1) maximum-index sorting; 2) Viterbi algorithm; 3) splitting.

Due to the lack of frame-level annotations, there is no annotation to locate the start frame and end frames per action. A frame can't know its correct corresponding sentence. To overcome this problem, we propose an essential hypothesis based on the temporal relation between sentences and frames: if s_j is the corresponding sentence representation for the frame representation h_i, the sentence representation for frame representation h_{i+1} should be in the set of $\{s_j, s_{j+1}, s_{j+2}, \ldots, s_K\}$ and never in the set of $\{s_1, s_2, \ldots, s_{j-1}\}$. The visualization of fine-grained contrastive loss can be seen in Fig. 3.

Specifically, in our model, we first obtain the sequence of frame-level sequences of representations H and sequence of sentence-level representations S through the vision-language module. As Eq. (7) shows, we symmetrically calculate the fine-grained contrastive learning loss, named L_{fine}, to achieve frame-sentence alignment.

$$\begin{aligned} L_{fine} = & \mathit{CE}(\psi_{\text{preds}}(H, S), \phi_{\text{pseudo}}(H, S)) \\ & + \mathit{CE}(\psi_{\text{preds}}(S, H), \phi_{\text{pseudo}}(S, H)) \end{aligned} \tag{7}$$

where *CE* is the Cross-Entropy loss. We use ψ_{preds} to predict the most related sentence s_j with one frame, where $s_j \in S$. The ϕ_{pseudo} could utilize the probability distribution of prediction and the similarity matrix of H and S to generate the pseudo labels as ground truth. Then as Eq. (7) shown, we calculate L_{fine} by the *CE* loss of the prediction and pseudo labels. And we separately introduce two methods of ψ_{preds} and ϕ_{pseudo} in Sec. 3.4.1 and Sec. 3.4.2. The method of Gumbel-Softmax with splitting is shown in Sec. 5.3.

3.4.1 Gumbel-Softmax with Sorting

We first use Eq. (5) to calculate the similarity matrix between the frame representations H and its sentence representations S. And we obtain the first prediction by Eq. (8).

$$\psi_{\text{preds}}(H, S) = \text{Gumbel-Softmax}(\varphi(H, S)) \tag{8}$$

where Gumbel-Softmax is the straight-through Gumbel-Softmax function [23]. We utilize the Gumbel-Softmax to ensure the dispersed sampling from the original distribution can be calculated for the gradients in the backward pass. Then, we get the maximum index through $\arg\max$ and sort the maximum-index list to an increasing order to generate pseudo labels. We regard them as the ground truth, which shows in Fig. 3b in blue. Finally, we finish the first kind of ψ_{preds} and ϕ_{pseudo} by Eqs. (8) and (9).

$$\phi_{\text{pseudo}}(H, S) = \text{sort}\left[\arg\max_{i \in [1,K]}(\psi_{\text{preds}}(H, S)_{N \times K})\right] \tag{9}$$

3.4.2 Gumbel-Softmax with Viterbi

Following the Viterbi algorithm [16], it could generate the maximum a posteriori probability estimate, called the Viterbi path. The original Viterbi algorithm needs two important matrices: transition matrix and emission matrix. As shown in Eq. (11), we use the similarity of the language and vision features as the emission with the shape $[N, K]$, where N means the number of sampled frames and K is the total number of its labels. Specifically in our method, as Eq. (10)shows, we use one upper triangular mask matrix as the transition matrix to limit the path of probability, which could make sure the way won't go back. Based on the Viterbi path (shown in Fig. 3c), we obtain the pseudo-labels by Eq. (12). Different from our method using Viterbi algorithm to generate pseudo labels, [24, 27] apply Viterbi decoding prediction, and activities have constant action orders. More details about the Viterbi algorithm can be seen in supplementary materials.

$$\text{Transition matrix:} A = \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ & \ddots & \vdots \\ 0 & & \frac{1}{n} \end{bmatrix}_{N \times N} \tag{10}$$

$$\text{Emission matrix: } B = \varphi(H, S) \tag{11}$$

$$\phi_{\text{pseudo}}(H, S) = \text{Viterbi}(A, B) \tag{12}$$

3.4.3 Training Loss

In conclusion, we train our module with the combination of the proposed coarse-grained contrastive loss and fine-grained contrastive loss:

$$L = L_{\text{coarse}} + \lambda_1 L_{\text{fine}} \tag{13}$$

where λ_1 represents the weight of fine-grained contrastive loss.

4. Experiments

In this section, we first introduce the implementation details, evaluation benchmarks and evaluation metrics in Sec. 4.1. The experiments to verify the effectiveness of baselines for video-text representation learning are shown in Sec. 4.2. In addition, we also transfer our proposed framework to downstream sequence verification in Sec. 4.3 and text-to-video matching tasks in Sec. 4.4.

4.1. Experimental Details

Implementation Details. The vision backbone we employ is the pre-trained CLIP vision encoder based on ViT-B [11]. And the model is initialized adopting Kaiming and Xavier uniform initialization for different layers [19, 20]. In our module, the parameter of the vision backbone is unfrozen and finetuned when training. On the other hand, the language backbone is the pre-trained CLIP text encoder whose parameter is frozen totally. More implementation details can be seen in supplementary materials.

Datasets. We conduct experiments on the datasets COIN-SV, Diving-SV and CSV. COIN-SV is rearranged from COIN and composed of 36 tasks that contain more than 20 comprehensive instructional videos in the training dataset. Diving-SV is rearranged from Diving and contains 48 kinds of diving competition videos. And CSV [41] includes 45 procedures for training and 25 procedures for testing. In these datasets, all kinds of videos in the test set are unseen in the training set.

Testing phase. During inference, we apply the method that distinguishes positive pairs from negative pairs to evaluate the quality of learned video representations. Specifically in this paper, we calculate the normalized Euclidean distance between two video representations v_1 and v_2 in the same video pair:

$$d = dis(v_1, v_2) \tag{14}$$

$$y = \begin{cases} 1, d \leq \tau \\ 0, otherwise \end{cases} \tag{15}$$

where $dis(.,.)$ means the $\ell 2$-normalization Euclidean distance function. τ is a threshold to decide whether the sequences are consistent. $y = 1$ means the two sequences of videos are consistent, otherwise inconsistent.

Evaluation Metrics. We adopt the Area Under ROC Curve (**AUC**) as the measurement for all of our experiments, which is commonly used to evaluate the performance in the field of anomaly detection [15] and face verification [10]. Higher AUC means better performance.

4.2. Comparison of baselines

Under weak supervision, the only annotations we know are the text descriptions of procedures, but the timestamps of actions and video task classification are unknown. The results of weakly supervised video sequence verification are shown in Tab. 1. We compare our method with other baselines, including 1) MIL-NCE [36]. 2) CAT, we change the SVIP [41] model architecture and add a text encoder to adapt to this task. 3) VideoSwin+MLP, we adopt the video swin transformer [33] as the vision encoder to extract frame features. 4) CLIP+Transformer Encoder+Pool. 5) CLIP+Transformer Encoder+MLP. To adapt to the task, we apply the CLIP text encoder as the text encoder of baselines except for MIL-NCE. Other methods but ours only calculate the coarse-grained contrastive loss.

Method	Text Encoder	Weakly Supervised (w/o CLS)		
		CSV	Diving-SV	COIN-SV
MIL-NCE [36]	MLP [36]	53.02	58.49	47.95
CAT [41]	CLIP [43]	70.63	77.87	47.70
VideoSwin [33]+MLP		62.48	60.88	**54.73**
CLIP [43]+TE [11]+Pool		58.67	72.13	49.79
CLIP [43]+TE [11]+MLP		74.82	81.47	50.13
Ours	CLIP [43]	**79.80**	**85.19**	52.56

Table 1. Results of representation learning for weakly supervised video sequence verification task.

The results in Tab. 1 demonstrate that multiple granularity contrastive learning is effective for learning discriminative video representations under weak supervision.

4.3. Sequence Verification

Following the setting of sequence verification [41], we know the classification of videos but yet do not know the

Method	Pre-train	Supervised (w CLS)		
		CSV	Diving-SV	COIN-SV
MIL-NCE [36]	HowTo100M [37]	56.16	63.43	47.80
Swin [32]	K-400 [5]	54.06	73.10	43.70
TRN [64]	K-400 [5]	80.32	80.69	57.19
CAT [41]	K-400 [5]	83.02	83.11	51.13
CLIP [43]+TE [11]+MLP	CLIP [43]	79.38	83.48	48.50
Ours (weakly supervised)	CLIP [43]	79.80	85.19	52.56
Ours	CLIP [43]	**86.92**	**86.09**	**59.57**

Table 2. Results of downstream video sequence verification task under supervised setting.

timestamp of actions. The testing results on sequence verification compared to other methods are shown in the Tab. 2. We can use class information for sequence verification of procedures in videos.

For a fair comparison, some adjustments have been made to the architecture of our model in this task setting. Specifically, we add a classification layer on the top of the video representation and the classification loss to our model. Besides, we apply the adjusted video sequence alignment mechanism by ours and train with pair data that are the same as SVIP [41]. This adjusted model is named ”Ours”. In addition, we also compare with some state-of-the-art methods [32, 41, 64] of sequence verification and change some video-language pre-trained model [37] accordingly to adapt to this task. Weakly supervised means no classification information of tasks. To clarify the improvements from technical differences, we replace the visual backbone of CAT [41] with CLIP-ViT [43] to form CLIP+TE+MLP. Then, we improve performance by adjusting network structure, e.g., the position of SEQ loss. Observed Tab. 2, our model outperforms them by a notable margin on all the considered datasets. The results also demonstrate that the fine-grained contrastive loss we proposed enforces the model to learn more discriminative representations. The results of our weakly supervised model, which surpasses other supervised methods, demonstrate our model's excellent performance.

4.4. Text-to-Video Matching

Setting. We validate the performance of the video-language representations on text-to-video matching, which aims to find the correct video corresponding to a sequence of texts from a series of videos. Specifically, we train our model on the CSV dataset under weak supervision and test it on our proposed benchmark about text-to-video matching. This task evaluates the model's ability to learn semantic and generalized video representations.

Benchmark. To better evaluate the text-to-video matching, we rearrange the test set of CSV [41] and propose a new scripted benchmark, named **CSV-Matching**. It has 800 text-video pairs. Each text-video pair is composed of one sequence of text descriptions of procedures and five videos. All of the videos describe the same task but hold different procedures. There is only one correct video matching the text descriptions in each pair. More details about the text-to-matching benchmark will show in the supplementary materials.

The text-to-video matching results in Tab. 3 indicate that our method has the best performance. And due to data of CSV-Matching being unseen when training, it shows that our method has a more powerful generalization ability.

Method	Text-to-Video Matching
	CSV-Matching
MIL-NCE [36]	60.02
CAT [41]	53.54
CLIP [43] +TE [11] +MLP	62.67
Ours	**65.23**

Table 3. Results of text-to-video matching task on our proposed benchmark *CSV-Matching*. We evaluate the results using AUC.

5. Analysis

In this section, we first analyze the impact of different backbones in Sec. 5.1. conduct comprehensive ablation studies of multiple granularity contrastive loss and pseudo-label generation in Secs. 5.2 and 5.3. Moreover, we analyze our limitations and broader impact.

5.1. Ablation of Backbone

As Tab. 4 shown, our method based on CLIP-ViT obtains the best performance compared with other backbones. In addition, results indicate that fine-grained and multi-grained losses improve performance under weak supervision and supervision, respectively.

Backbone	Pretrained	Weakly Supervised (w/o CLS)			Supervised (w CLS)	
		L_{coarse}	L_{fine}	CSV	$L_{\text{coarse}}+L_{\text{fine}}$	CSV
ResNet50 [21]	ImageNet-1K	✓	✗	76.22	✗	78.83
			✓	78.32	✓	81.00
ViT-B/32 [11]	ImageNet-21K	✓	✗	73.88	✗	81.66
			✓	75.18	✓	82.11
CLIP-ViT [43] (Ours)	Text-Image Pair	✓	✗	78.49	✗	83.58
			✓	79.80	✓	86.92

Table 4. Results of our method with different backbone on CSV.

5.2. Ablation of Multiple Granularity Contrastive Loss

In this section, we conduct comprehensive ablation studies to investigate the effects of our multiple granularity contrastive loss. To better demonstrate the superiority of our method, we present the loss ablation experiments on the sequence verification task under supervision with classification in Tab. 5. As shown, both coarse-grained contrastive loss L_{coarse} and fine-grained loss L_{fine} are crucial. Specifically, the method with coarse-grained and fine-grained contrastive loss surpasses the method without them by 3.34%.

Method	L_{fine}	L_{coarse}	CSV
Ours (w CLS)	✗	✗	83.58
	✓	✗	84.85
	✗	✓	84.32
	✓	✓	**86.92**

Table 5. Ablation studies of our proposed multiple granularity contrastive loss on CSV. To verify the effectiveness of L_{fine} and L_{coarse} separately, we conduct experiments on video verification task.

Introducing the fine-grained loss L_{fine} brings 2.6% performance improvement compared to only using coarse-grained contrastive loss L_{coarse}. Comparing only uses L_{coarse} or uses L_{fine}, the result indicates that the model training with more fine-grained information is better than coarse information. By restricting the video representation to frame-sentence level latent space, the fine-grained contrastive loss can help the model learn more discriminative video representations.

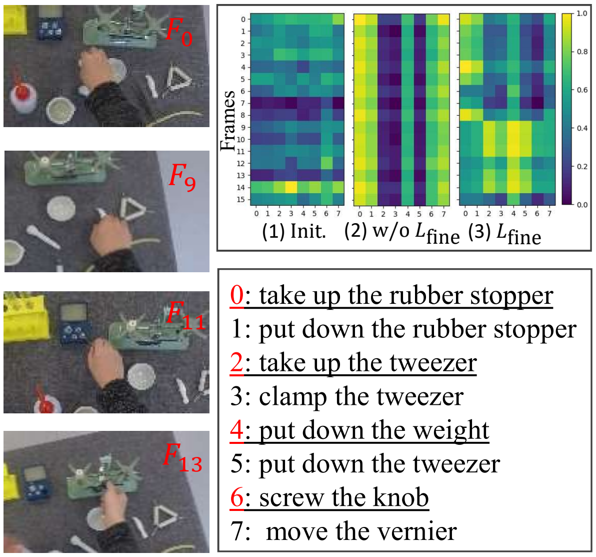


Figure 4. **Visualization** of ablation study about fine-grained contrastive loss.

The visualization of the ablation study about fine-grained loss, as shown in Fig. 4, illustrates fine-grained contrastive loss implements the alignment between frames and sentences.

5.3. Ablation of the Pseudo-Label Generation

Splitting. Splitting means we split the sequence of frame representations or sentence representations uniformly into several parts to keep the sequence length of frame representations or sentence representations equal. The values belonging to the same part will be added and then averaged. After that, we get a square matrix and output probability distribution of prediction. The elements along the diagonal are regarded as pseudo labels. Then calculate the fine-grained contrastive loss as Eq. (7). This process is shown in Fig. 3d, and the blue boxes represent the pseudo-labels.

Method	L_{fine}	Pseudo-label generation	CSV
Ours	✗	✗	74.82
	✓	split	72.75
		viterbi	78.46
		sort	**79.80**

Table 6. Ablation studies of the type of pseudo-label generation on our proposed method.

We conduct ablation studies about three methods of pseudo-label generation in the fine-grained loss L_{fine} showing in Tab. 6. Specifically, we validate the effectiveness of different kinds of coarse-grained contrastive loss on the weakly supervised video verification task. The results show the algorithms of maximum-index sorting and Viterbi are performing better than splitting. The method of splitting matrices into several parts and aligning sequences along the diagonal is too simple and crude .

Broader Impact and Limitations. In realistic sequential videos, sub-actions could be repeated. It could mislead the model to generate biased pseudo-labels and lead to the deterioration of performance.More analysis can be seen in supplementary materials. Moreover, the proposed method will likely be applied to behavior detection, healthcare, online education, industrial generation, etc.

6. Conclusions

In this paper, we propose a novel framework of weakly supervised video representation learning for sequential videos. Borrowing the multi-modal contrastive learning from CLIP, our method can learn video representation with unaligned text and video without relying on the accurate time-stamp level text-video annotation. We propose a multiple granularity loss where the video-paragraph contrastive loss constrains the matching between the whole video and the complete script, and a fine-grained frame-sentence contrastive loss constrains the matching between each action and its descriptions. We also propose to generate pseudo labels with temporal consistency in video and text. Experiments results show that our design is effective, and our method achieves state-of-the-art performance when transferred to downstream video sequence verification and text-to-video matching tasks.

Acknowledgements.

The work was supported by National Key R&D Program of China (2018AAA0100704), NSFC #61932020, #62172279, Science and Technology Commission of Shanghai Municipality (Grant No. 20ZR1436000), and “Shuguang Program” supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission.

Weakly Supervised Video Representation Learning with Unaligned Text for Sequential Videos

Supplementary Material

A. Extra Experiment Studies

In this section, we present additional ablation studies about our method, including the effects of batch size, the number of clips sampled per video, the approach of extracting paragraph-level language representation, and sequence align loss.

A.1. Implementation Details

We implement our method with PyTorch. The vision backbone we employ is the pre-trained CLIP vision encoder based on ViT-B [11]. And the model is initialized adopting Kaiming and Xavier uniform initialization for different layers [19, 20]. In our module, the parameter of the vision backbone is unfrozen and finetuned when training. On the other hand, the language backbone is the pre-trained CLIP text encoder whose parameter is frozen totally. We split the raw video into 16 clips for a sequential video and randomly sample one raw frame from each clip in the training and uniformly sample frames in inference. The projection layer adopts a fully connected layer. The hidden layer dimension of transformer encoder [11] is 1024, and the depth is 2. The dimension of the video representations and paragraph representations is 512. The λ_1 in our model is equal to 1. The experiments are conducted on 4 NVIDIA 2080Ti GPUs with batch size 8. We adopt an AdamW optimizer [60] with cosine annealing learning rate scheduler with a base learning rate of 5×10^{-4}, and weight decay 0.01. More implementation details should be seen in the supplementary materials. And expect the experiment of sequence align loss to be conducted on the supervised sequence verification task, other experiments are conducted on the weakly supervised sequence verification task. We conduct all experiments on CSV dataset.

A.2. Batch size

To adapt to the change in batch size, we increase or decrease the learning rate exponentially. As Tab. 7 shown, our method achieves the best performance when the batch size is equal to 16. The larger the batch size, the more likely multiple videos of the same task will appear in the same mini-batch. Due to the limitation of GPU memory, the largest batch size can be set as 8 if we unfreeze the vision backbone.

Method	Batch size	Frames	CSV
Ours	4	16	65.94
	8	16	67.46
	16	16	69.42
	24	16	69.21
	32	16	69.16

Table 7. Ablation studies of batch size on our proposed method

A.3. Sampling

While changing the frames of sampling per video, the training time is doubled with the increasing number of frames. In this ablation study, all models have been training no more than 100 epochs or 12 hours on two GPUs due to the limitation of computing resources.

As Tab. 8 shown, when frames are set to 16, our model achieves the best performance. It is worth noting that, with more training steps (about twice the training time), the performance of 32 frames will increase to 68.20. However, we choose 16 as the default frame to balance batch size, number of frames, and training cost, we choose 16 as the default frame.

The significant reason for choosing sampling frames rather than video clips is the limitation of computational resources. Fine-tuning the full pre-trained backbone, such as VideoCLIP [59], is expensive. Similarly, to balance the efficiency of the network and fairly compare our method with CAT [41], we choose 16 frames as the same as CAT.

Method	Batch size	Frames	CSV
Ours	8	8	58.43
	8	16	67.46
	8	32	65.64
	8	48	64.40

Table 8. Ablation studies of the number of frames.

A.4. Paragraph feature

We design two ways to extract the feature of the paragraph. The one is concatenating all sentences into a paragraph description. Then we can obtain the paragraph-level representation by feeding the paragraph description into the language encoder. The other method is that feed individual procedure texts into the frozen language encoder to produce sentence representations and then obtain a paragraph-level representation by temporal mean pooling. The results shown in Tab. 9 illustrate that the method based on concatenation achieves better performance.

Method	Paragraph feature	CSV
Ours	pooling	67.07
	concat	67.46

Table 9. Ablation studies of the ways to extract paragraph features on our method.

A.5. Sequence alignment loss

For a fair comparison, some adjustments have been made to the architecture of our model on the supervised sequence verification task. Specifically, following [41], we apply the video sequence alignment mechanism to our model. Moreover, we also conduct experiments to investigate the effectiveness of using sequence alignment loss. We change the sequence align loss position to the last of the network. The results shown in Tab. 10 illustrate that sequence alignment loss L_{seq} could restrict the model to learning a better representation.

Method	L_{seq}	CSV
Ours	✗	84.47
	✓	84.69

Table 10. Ablation studies of the sequence alignment loss on our method.

B. Gumbel-Softmax with Viterbi

Due to the sum of the probabilities of each row cannot be greater than one and each probability value in a row should be the same, we simply set the value to $\frac{1}{N}$. As Eq. (16) shown, we set each element value in the upper diagonal matrix to $\frac{1}{N}$ and others to zero to keep the path of probability will be a one-way path.

$$A = \begin{bmatrix} \frac{1}{N} & \cdots & \frac{1}{N} \\ & \ddots & \vdots \\ 0 & & \frac{1}{N} \end{bmatrix}_{N \times N} \tag{16}$$

where A represents the Transition matrix of Viterbi algorithm [16].

C. TSM module

Following [12], we add the Temporal Similarity Matrix (TSM) module with residual connection to our vision module. In this ablation study, we only use the task classification loss L_{cls} instead of coarse-grained loss L_{coarse} and fine-grained loss L_{fine}. As Tab. 11 shown, we verify different similarity distances of TSM and residual connection types. And the experiments indicate that the TSM module with residual connection will improve the model performance.

However, as Tab. 12 shows, while we apply the TSM module to our method and train the model under weak supervision, the performance of the model degrades. It is reasonable that the model with the TSM module is not effective for language-video alignment tasks.

Method	Dist	Residual	CSV
CLIP [43]+TE [11]+MLP	✗	✗	77.35
	L2	add	77.42
	L2	concat	78.22
	Attn	add	76.89
	Attn	concat	77.71

Table 11. Ablation studies of the different kinds of TSM module on the baseline.

Method	L_{fine}	L_{coarse}	TSM	CSV
Ours	✓	✓	✗	79.80
	✓	✓	✓	76.00

Table 12. Ablation studies of the TSM on our proposed method.

D. Downstream tasks

D.1. Text-to-Video Matching

We validate the performance of the video-language representations on text-to-video matching, which aims to find the correct video corresponding to a sequence of texts from a series of videos. Specifically, we train our model on the CSV dataset under weak supervision and test it on our proposed benchmark about text-to-video matching. We calculate the similarity between each video representation V_i and paragraph representation L:

$$d = dis(L, V_i) \tag{17}$$

where $dis(.,.)$ represents the normalized Euclidean distance. And V_i represents $i_{th} (i \in [0, \ldots, 4])$ video representation. At last, we select the text-video pair with the max similarity.

CSV-Matching. To better evaluate the text-to-video matching, we rearrange the test set of CSV and propose a new scripted benchmark named CSV-Matching. It has 800 text-video pairs. Each text-video pair is composed of one sequence of text descriptions of procedures and five videos. All of the videos describe the same task but hold different procedures. There is only one correct video matching the text descriptions in each pair. CSV-test dataset contains 5 tasks and each task has 5 kinds of different procedures. We random select one kind of video from each procedure to compose one pairs. The benchmark and split script will be available.

Method	Backbone	Loss	Classification(Acc)
CAT [41]	ResNet-50 [12]	CLS, SEQ	61.08
CLIP [43]+TE+MLP	CLIP-ViT	CLS, SEQ	63.24
Ours(CLS)	CLIP-ViT	CLS, SEQ, Multi-grained loss	**69.57**

Table 13. Results of video classification on CSV.

Method	Backbone	Weakly supervised (w/o CLS)			Supervised (w CLS)		
		Def.	No Rep.	Rep.	Def.	No Rep.	Rep.
CAT [41]	ResNet50 [21]	47.70	57.82	49.99	51.13	63.25	45.96
CLIP [43]+TE+MLP	CLIP-ViT	50.83	65.28	53.73	48.50	65.21	51.25
Ours	CLIP-ViT	**52.55**	**68.98**	**56.16**	**59.57**	**77.78**	**54.95**

Table 14. Results of different methods on re-divided COIN-SV.

D.2. Video Classification

To demonstrate our method's transfer ability, we evaluate models in the downstream video classification task. We re-divided the CSV dataset for video classification task. The train set contains 689 videos and test set contains 185 videos. On the re-divided CSV test dataset (CSV-CLS), we evaluate representations of models with linear probing, which were pre-trained under weak supervision. As Tab. 13 shown, our method achieves better performance in the video classification task. The benchmark and split script will be available.

E. Limitations

While our method performs well on the major part of the data, there still are some failure cases. In realistic sequential videos, sub-actions are often repeated. In that case, there are multiple sentences with high similarity to a frame. It could mislead the model to generate biased pseudo-labels, which will lead to the deterioration of performance. For example, the occurrence of a large number of repetitive actions repetitive action might hidden achieving further performance.

The intuition of our fine-grained contrastive loss comes from a basic idea: *if the s_j is the corresponding sentence for frame h_i, the corresponding sentence for frame h_{i+1} is never before the s_j in sequence*. Due to a large number of repetitive actions, it might be difficult to achieve further performance. However, this method is still promising. As Tab. 14 shown, we have re-divided the COIN-SV test dataset based on whether existing repetitive actions in videos or not, which are COIN-SV-Rep (675 video pairs) and COIN-SV-NoRep (325 video pairs). In the original COIN-SV test dataset, there are 1000 video pairs for sequential video verification, built by 328 videos containing repetitive actions and 123 videos that do not. The results show that although the occurrence of repetitive actions will cause the deterioration of performance, our method can still achieve better results than other baselines. Moreover, the results conducted by our method may reflect the bias from the dataset.

References

[1] Nadine Behrmann, S Alireza Golestaneh, Zico Kolter, Juergen Gall, and Mehdi Noroozi. Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In *European Conference on Computer Vision*, pages 52–68. Springer, 2022. 1, 2

[2] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 3

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3

[4] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. *arXiv preprint arXiv:2207.10362*, 2022. 1, 3

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 3, 7

[6] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3546–3555, 2019. 3

[7] Minghao Chen, Fangyun Wei, Chong Li, and Deng Cai. Frame-wise action representations for long videos via sequence contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13801–13810, 2022. 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning

of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 3

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[11] 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. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4, 6, 7, 9, 10

[12] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10387–10396, 2020. 10, 11

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 3

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3

[15] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021. 6

[16] G David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973. 2, 5, 10

[17] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridgeformer: Bridging video-text retrieval with multiple choice questions. *arXiv preprint arXiv:2201.04850*, 2022. 1, 3

[18] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2906–2916, 2022. 1, 3

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 6, 9

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 6, 9

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7, 11

[22] Huazhang Hu, Sixun Dong, Yiqun Zhao, Dongze Lian, Zhengxin Li, and Shenghua Gao. Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19013–19022, 2022. 3

[23] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 2, 5

[24] Anna Kukleva, Hilde Kuehne, Fadime Sener, and Jurgen Gall. Unsupervised learning of action classes with continuous temporal embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12066–12074, 2019. 6

[25] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M Zeeshan Zia, and Quoc-Huy Tran. Unsupervised action segmentation by joint representation learning and online clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20174–20185, 2022. 3

[26] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017. 1

[27] Jun Li and Sinisa Todorovic. Action shuffle alternating learning for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12628–12636, 2021. 6

[28] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429, 2022. 3

[29] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 2

[30] Chung-Ching Lin, Kevin Lin, Lijuan Wang, Zicheng Liu, and Linjie Li. Cross-modal representation learning for zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19978–19988, 2022. 3

[31] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863, 2022. 3

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10012–10022, 2021. 3, 7

[33] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 3, 6

[34] Zijia Lu and Ehsan Elhamifar. Set-supervised action learning in procedural task videos via pairwise order consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19903–19913, 2022. 3

[35] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 3

[36] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 6, 7

[37] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 3, 7

[38] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009. 3

[39] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 3

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[41] Yicheng Qian, Weixin Luo, Dongze Lian, Xu Tang, Peilin Zhao, and Shenghua Gao. Svip: Sequence verification for procedures in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19890–19902, 2022. 3, 6, 7, 9, 10, 11

[42] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021. 1

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3, 4, 6, 7, 10, 11

[44] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 3

[45] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020. 3

[46] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 3

[47] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 3

[48] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pretraining with multimodal temporal contrastive learning. *arXiv preprint arXiv:2210.06031*, 2022. 1, 3

[49] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 2

[50] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1

[51] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3

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

[53] Jue Wang, Gedas Bertasius, Du Tran, and Lorenzo Torresani. Long-short temporal contrastive learning of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14010–14020, 2022. 3

[54] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1

[55] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 3

[56] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment

networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 1

[57] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3

[58] 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. 1

[59] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 1, 3, 9

[60] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019. 9

[61] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1

[62] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762, 2015. 3

[63] Yanan Zhang, Jiaxin Chen, and Di Huang. Cat-det: Contrastively augmented transformer for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 908–917, 2022. 1

[64] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018. 7

[65] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 1, 3