LocVTP: Video-Text Pre-training for Temporal Localization

Meng Cao[1*], Tianyu Yang[2‡], Junwu Weng[2], Can Zhang[1], Jue Wang[2], and Yuexian Zou[1,3†]

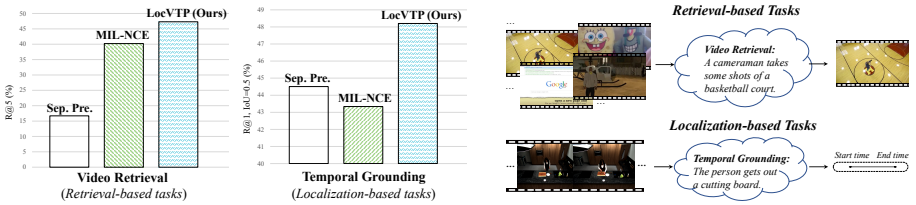
[1] School of Electronic and Computer Engineering, Peking University
[2] Tencent AI Lab
[3] Peng Cheng Laboratory

Abstract. Video-Text Pre-training (VTP) aims to learn transferable representations for various downstream tasks from large-scale web videos. To date, almost all existing VTP methods are limited to *retrieval-based* downstream tasks, *e.g.*, video retrieval, whereas their transfer potentials on *localization-based* tasks, *e.g.*, temporal grounding, are underexplored. In this paper, we experimentally analyze and demonstrate the incompatibility of current VTP methods with localization tasks, and propose a novel **Loc**alization-oriented **V**ideo-**T**ext **P**re-training framework, dubbed as **LocVTP**. Specifically, we perform the fine-grained contrastive alignment as a complement to the coarse-grained one by a clip-word correspondence discovery scheme. To further enhance the temporal reasoning ability of the learned feature, we propose a context projection head and a temporal aware contrastive loss to perceive the contextual relationships. Extensive experiments on four downstream tasks across six datasets demonstrate that our LocVTP achieves state-of-the-art performance on both retrieval-based and localization-based tasks. Furthermore, we conduct comprehensive ablation studies and thorough analyses to explore the optimum model designs and training strategies. Codes are available at https://github.com/mengcaopku/LocVTP.

1 Introduction

Video-Text Pre-training (VTP) [49,39,31,30,26,4,67,54] has attracted increasing attention with the aim to learn generic and transferable *joint* video-language (VL) representations. Compared to the conventional *separate* pre-training on each single modality, *e.g.*, video features are pre-trained under the action recognition datasets (Kinetics [23], Sport1M [22]), VTP has several advantages: 1) It leverages large-scale unlabeled narrated video data with automatically generated corresponding text data for video-text correspondence pre-training. 2) It tries to map different modality features into a shared latent space, which reduces the difficulties of the cross-modal feature interaction. Thanks to these advantages, VTP has significantly improved the performance of many downstream VL tasks.

* Work done during an internship at Tencent AI Lab.
‡ project leader. † corresponding author.



(a) (b)

Fig. 1: (a) **Video retrieval and temporal grounding performance using different pre-trained features.** `Sep.Pre.` means separately pre-training, *i.e.*, the video encoder supervisedly pre-trained on Kinetics [23] and text encoder taken from BERT [12]. `MIL-NCE` and our `LocVTP` are VTP methods pre-trained on HowTo100M [39]. For video retrieval, we use COOT [16] as the downstream method and evaluate on YouCook2 [75] dataset with R@5. For temporal grounding, we take 2D-TAN [74] as the downstream method and evaluate on ActivityNet Captions [24] dataset with R@1, IoU=0.5. **(b) Retrieval-based and localization-based downstream tasks.** We take video retrieval and temporal grounding as typical examples, respectively. The former needs video-level classification while the latter requires clip-level or frame-level localization.

For example, as illustrated in [16], the video retrieval performance using features pre-trained with the VTP method `MIL-NCE` [38] is much higher than that using separately pre-trained way (cf. Fig. 1a (left)).

Despite their encouraging performance, we find that most current VTP methods are applicable to limited downstream tasks, *i.e.*, they focus on *retrieval-based* tasks which require video-level predictions, *e.g.*, video retrieval [64], video captioning [45], and video question answering [21]. In contrast, there exists another mainstream *localization-based* tasks which expect more fine-grained clip-level or frame-level predictions, *e.g.*, temporal grounding [15], action segmentation [51], action step localization [77] (cf. Fig. 1b). Unfortunately, through experiments, we find their poor generalization abilities on this type of downstream tasks. For example, on temporal grounding, even pre-trained with a much larger dataset HowTo100M [39], the VTP method `MIL-NCE` still performs worse than the separately pre-trained counterpart (cf. Fig 1a (right)).

In this paper, we analyze that this poor transfer ability on localization-based tasks is due to the absence of two indispensable characteristics: **1) *Fine-grained alignment***: We contend that the alignment should be conducted on more *fine-grained* clip-word level instead of the *coarse-grained* video-sentence[4] level. As the temporal grounding example shown in Fig. 2, a given query sentence may contain multiple actions (*e.g.*, "`hit the golf ball`" (q^{s1}) and "`bend down to pick up the ball`" (q^{s2})). Thus, aligning each action (or words) to the corresponding clips (*i.e.*, v^{t1} and v^{t2}) will help to obtain more detailed and accurate feature representations. **2) *Temporal relation reasoning***: We hope the clip features of a certain action can also perceive other actions in the same video. For example, for a typical `golf` video, action q^{s2} ("`bend down to pick up the`

[4] Here we use "sentence" to represent the whole paired text for each video, such as the ASR in HowTo100M [39] or query language in ActivityNet Caption [24].

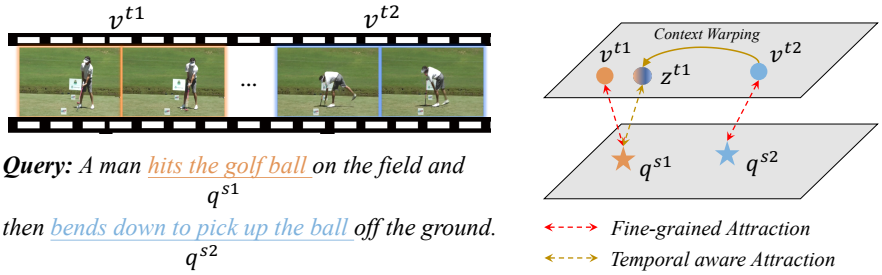


Fig. 2: **Fine-grained video-text alignment:** positive clip-word pairs are selected via cosine similarity and are then forced to be close to each other, *i.e.*, $\boldsymbol{v}^{t1} \leftrightarrow \boldsymbol{q}^{s1}$, $\boldsymbol{v}^{t2} \leftrightarrow \boldsymbol{q}^{s2}$; **Temporal relation reasoning:** a context warping head reconstructs $\boldsymbol{v}^{t1}$ conditioned on $\boldsymbol{v}^{t2}$ and distance $t2 - t1$ while maintaining the cross-modal alignment unchanged, *i.e.*, $\boldsymbol{z}^{t1} = \text{warp}(\boldsymbol{v}^{t2}, t2 - t1) \leftrightarrow \boldsymbol{q}^{s1}$.

`ball`") always occurs shortly after action q^{s1} ("`hit the golf ball`"). Thus, incorporating such temporal relationship into VTP can help to improve the temporal awareness of video features.

Based on these observations, we propose a novel video-text pre-training framework for localization tasks, dubbed as LocVTP. By considering both above-mentioned characteristics, LocVTP achieves state-of-the-art performance not only on the widely studied retrieval-based tasks, but also on the less-focused localization-based tasks. Specifically, **for fine-grained alignment**, we extend the coarse-grained contrastive training with video-sentence alignment to a fine-grained one with clip-word alignment. Since there are no clip-word correspondence annotations in existing large-scale datasets, we utilize the latent space established by the coarse-grained contrastive learning to estimate the clip-word similarity, and then select the clip-word pairs with high similarities as positive samples. To further illustrate this, as shown in Fig. 2 (right), suppose $\{\boldsymbol{v}^{t1}, \boldsymbol{q}^{s1}\}$ and $\{\boldsymbol{v}^{t2}, \boldsymbol{q}^{s2}\}$ are two matched clip-word feature pairs. Semantic embeddings in each pair are mapped to be close to each other, *i.e.*, $\boldsymbol{v}^{t1} \leftrightarrow \boldsymbol{q}^{s1}$, $\boldsymbol{v}^{t2} \leftrightarrow \boldsymbol{q}^{s2}$. **For temporal relation reasoning**, we propose a new pretext task called *context warping*. Here we use Fig. 2 (right) for illustration. Context warping is designed to generate a new temporally relevant clip features $\boldsymbol{z}^{t1}$, which imitates $\boldsymbol{v}^{t1}$, conditioned on another clip $\boldsymbol{v}^{t2}$ and the relative distance $t2 - t1$ in time, *i.e.*, $\boldsymbol{z}^{t1} = \text{warp}(\boldsymbol{v}^{t2}, t2 - t1)$. The predicted relevant clip feature $\boldsymbol{z}^{t1}$ is enforced to maintain the original established cross-modal correspondence unchanged, *i.e.*, $\boldsymbol{z}^{t1} \leftrightarrow \boldsymbol{q}^{s1}$. In this manner, we simulate the contextual reasoning process and enhance the temporal awareness of video features.

We conduct extensive experiments on four downstream tasks (*i.e.*, video retrieval, temporal grounding, action step localization, and action segmentation) across six datasets. The results on both retrieval-based and localization-based tasks demonstrate the superiority and the generalization ability of our LocVTP.

In summary, we make three contributions in this paper:

- We propose a localization-oriented video-text pre-training framework, LocVTP, which benefits both retrieval-based and the less-explored localization-based downstream tasks.
- We pinpoint two crucial designs in LocVTP, *i.e.*, fine-grained video-text alignment and temporal relation reasoning.
- Experimental results show that our LocVTP significantly outperforms previous state-of-the-art methods when transferred to various downstream tasks.

2 Related Work

Video-Text Pre-training (VTP). With the release of the large-scale instructional dataset HowTo100M, VTP has spurred significant interest in the community. Overall, the mainstream methods can be broadly classified into two classes: 1) Generative methods: Several methods [28,34,50,11,20,56,31,55] try to extend BERT [53] to the cross-modal domain, *i.e.*, they accept both visual and textual tokens as input and perform the masked-token prediction task. 2) Discriminative methods. These methods [26,4,41,30] learn representations by differentiating input samples using objectives such as the metric loss [19,58] or contrastive loss [18,9]. ClipBert [26] enables affordable pre-training from sparsely sampled frames. Frozen [4] adapts the recent ViT [13] as the visual encoder and is flexible to be trained on both image and video datasets. T2VLAD [56] and FCA [17] also perform the fine-grained interactions between video clips and phrases. However, both of them resort to additional overload, *e.g.*, k-means cluster or graph auto-encoder. In contrast, our LocVTP explicitly models the clip-word matching with a more light-weighted similarity comparison manner.
Pre-training for localization tasks. Compared to the retrieval tasks [64,45,21] which only require only video-level predictions, localization tasks [15,51,77] are essentially different since they need dense clip-level or frame-level predictions and thus the pre-training for these tasks is more challenging. In the pure video domain, this gap has been noticed and several pre-training works [65,2,66,73] tailored for action localization have been proposed. BSP [65] synthesizes temporal boundaries using existing action recognition datasets and conducts boundary type classification to generate localization-friendly features. TSP [2] trains video encoders to be temporally sensitive by predicting the foreground clip label and classifying whether a clip is inside or outside the action. As for the video-language domain, our LocVTP is the first pre-training framework designed for localization tasks. Besides, compared to TSP and BSP which require label information for supervised pre-training, our LocVTP can directly learn from narrated videos.

3 Approach

3.1 Overview of LocVTP

An overview of LocVTP is illustrated in Fig. 3. We firstly feed the video and language modalities to their respective encoders $f_v(\cdot)$ and $f_q(\cdot)$ to obtain embedded features. We follow the sparse sampling spirit in [26] and sample T clips

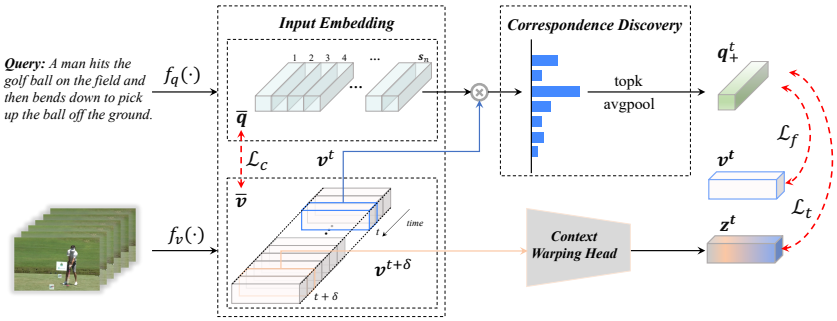


Fig. 3: An overview of LocVTP. $f_v(\cdot)$ and $f_q(\cdot)$ are video and language encoders, respectively. **1)** Coarse-grained contrastive loss $\mathcal{L}_c$ matches the global video and sentence representations $\overline{\boldsymbol{v}}$ and $\overline{\boldsymbol{q}}$. **2)** The clip-word correspondence is firstly built by similarity computing and then fine-grained contrastive loss $\mathcal{L}_f$ conducts detailed cross-modal alignment. Note that for clarity, we only present the correspondence discovery for the clip $\boldsymbol{v}^t$. **3)** A context warping head is employed to warp the contextual feature $\boldsymbol{v}^{t+\delta}$ and a temporal aware contrastive loss $\mathcal{L}_t$ is applied based on the warped feature $\boldsymbol{z}^t$.

for each video, yielding the encoded video $\boldsymbol{v} = \{\boldsymbol{v}^t\}_{t=1}^{T}$, where $\boldsymbol{v}^t \in \mathbb{R}^D$ is the t^{th} clip feature and D is the feature dimension. The text embedding is represented as $\boldsymbol{q} = \{\boldsymbol{q}^s\}_{s=1}^{S_q}$, where $\boldsymbol{q}^s \in \mathbb{R}^D$ is the s^{th} word embedding and S_q is the word length of $\boldsymbol{q}$.

Three types of contrastive methods are then performed to learn cross-modal features: 1) The coarse-grained contrastive loss builds the video-sentence level alignment; 2) A correspondence discovery strategy is proposed to build clip-word relations, based on which the fine-grained contrastive loss is applied; 3) Temporal aware contrastive loss with the context warping pretext task is proposed to encode temporal information into video representations.

3.2 Coarse-grained Contrastive Learning

We firstly conduct contrastive alignment at the global video-sentence level. Specifically, to obtain the video and sentence level features, we average pool $\boldsymbol{v}$ and $\boldsymbol{q}$ along the temporal and word index dimension, respectively. The global features are represented as $\overline{\boldsymbol{v}}$, $\overline{\boldsymbol{q}} \in \mathbb{R}^D$. Then we formulate this video-sentence alignment into the contrastive framework [18] as follows:

$$\mathcal{L}_c = -\log \frac{\exp\left(\overline{\boldsymbol{v}} \cdot \overline{\boldsymbol{q}} / \tau\right)}{\sum_{i=1}^{N} \exp\left(\overline{\boldsymbol{v}} \cdot \overline{\boldsymbol{q}}_i / \tau\right)}, \tag{1}$$

where $\overline{\boldsymbol{q}}_i, i \in [1, N]$, is the sentence feature for other samples within the batch. N denotes the batch size and τ is the temperature parameter. The coarse-grained contrastive loss $\mathcal{L}_c$ serves as a base loss to conduct video-sentence level constraint and induces a basic latent space where the detailed cross-modal matching is achieved. Though usually coarse and noisy, this latent space encodes prior for fine-grained clip-word correspondence discovery. In Section 4.6, we design and analyze three potential ways to use this cross-modal matching prior.

3.3 Fine-grained Contrastive Learning

Beyond the coarse-grained video-sentence alignment, we propose to conduct contrastive learning in a fine-grained manner, *i.e.*, clip-word matching. We contend that introducing such alignment learning into the pre-training stage could narrow down its gap with downstream localization tasks and calibrate the pre-trained feature to be more temporally aware.

Clip-word correspondence discovery. Before performing fine-grained contrastive learning, we firstly need to estimate the clip-word correspondences from video-sentence pairs. Thanks to the priors well established by the coarse-grained contrastive learning, we compute the cosine similarities between the video clips and their corresponding caption words in the pre-built latent space and choose the most similar K words as the correspondence for each video clip. Note that we select multiple positive words rather than simply pick one with the highest similarity because individual words may have vague meanings while sense-group[5] conveys more precise information (cf. Section 4.7).

Given the video sentence pair $\{\boldsymbol{v}, \boldsymbol{q}\}$, for the encoded t^{th} video clip $\boldsymbol{v}^t$, we compute its cosine similarities with the s^{th} word embedding $\boldsymbol{q}^s$ and apply the topk operation to select the most matched K ones. Following [57], these K selected items are average pooled to form the final positive sample:

$$\boldsymbol{q}_+^t = \text{avgpool}\left(\underset{s\in[1,S_q]}{\arg\text{topk}}\left(\boldsymbol{v}^t \cdot \boldsymbol{q}^s\right)\right), \tag{2}$$

where $\boldsymbol{q}_+^t$ is the final positive sample for $\boldsymbol{v}^t$. $(\boldsymbol{u}\cdot\boldsymbol{v}) = \boldsymbol{u}^\top\boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$ represents the cosine similarity between ℓ_2 normalized $\boldsymbol{u}$ and $\boldsymbol{v}$. This process can be efficiently performed for all the video clips using matrix operations.

Fine-grained contrastive loss. With the selected clip-word correspondence as positive pairs, we perform fine-grained representation learning following the cross-modal InfoNCE [18] loss (cf. Figure 4a). The negative samples are taken from the other words within the batch. Therefore, the fine-grained contrastive loss is defined as follows.

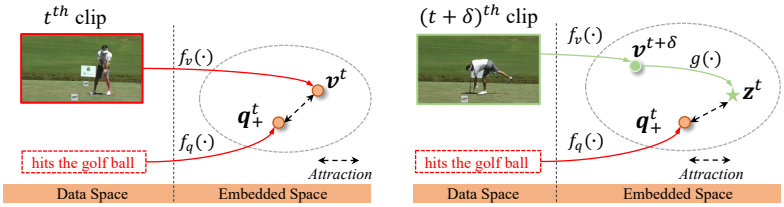
$$\mathcal{L}_f = \frac{1}{T}\sum_{t=1}^{T} -\log\frac{\exp(\boldsymbol{v}^t\cdot\boldsymbol{q}_+^t/\tau)}{\sum_{i=1}^{N}\sum_{s=1}^{S_{q_i}}\exp\left(\boldsymbol{v}^t\cdot\boldsymbol{q}_i^s/\tau\right)}, \tag{3}$$

where $\boldsymbol{q}_i^s$ is the s^{th} word feature of the i^{th} sentence $\boldsymbol{q}_i$.

3.4 Temporal aware Contrastive Learning

Compared with the video-level retrieval task, which favors temporal invariant features [40,42], the clip-level localization task [8,33,61,6,60,70,71,7,72] prefers temporal aware video embeddings. Specifically, correlated actions in the same video should perceive each other. This characteristic is however not embodied in the aforementioned contrastive learning.

[5] A group or sequence of words conveying a particular meaning or idea in linguistics.



(a) (b)

Fig. 4: Illustrations of (a) fine-grained contrastive Loss and (b) temporal aware contrastive loss. $\boldsymbol{v}^t$ is the t^{th} clip of video $\boldsymbol{v}$. $\boldsymbol{q}_+^t$ is the pooled positive word features. $\boldsymbol{z}^t$ is the warped feature. We only present positive samples and omit negative ones.

Context warping head. To alleviate this, we set up a *context-warping* operation to enforce the video clip to perceive the context. For the video clip $\boldsymbol{v}^t$ in a matched clip-word pair $\{\boldsymbol{v}^t, \boldsymbol{q}_+^t\}$ (cf. Section 3.3), we warp its contextual video clip with δ temporal distance, *i.e.*, $\boldsymbol{v}^{t+\delta}$, to "reconstruct" itself. To supervise this warping process, we set up a temporal aware contrastive loss to maintain the established correspondence. Specifically, we propose a *context warping head* $g(\cdot)$ to instantiate this warping process, by taking the context clip feature $\boldsymbol{v}^{t+\delta}$ and temporal distance δ as input.

$$\begin{aligned} \boldsymbol{z}^t &= g(\boldsymbol{v}^{t+\delta}; \delta) \\ &= \text{ReLU}\left(W[\boldsymbol{v}^{t+\delta}, \text{sgn}(\delta), |\delta|]\right), \end{aligned} \tag{4}$$

where $\boldsymbol{z}^t$ is the warped feature. $W \in \mathbb{R}^{(D+2)\times D}$ are the trainable weights. δ is randomly sampled within the range of $[-\delta_{max}, \delta_{max}]$. $\text{sgn}(\cdot)$ is the sign function which returns 1 for positive values and -1 for negative ones. Here $\text{sgn}(\delta)$ and $|\delta|$ indicate the direction and distance of the temporal difference δ, respectively.

Temporal aware contrastive loss. Through the context warping head, the warped feature $\boldsymbol{z}^t$ should mimic the reference feature $\boldsymbol{v}^t$. Since $\boldsymbol{v}^t$ has the clip-word alignment with $\boldsymbol{q}_+^t$, such correspondence should be preserved between the warped feature $\boldsymbol{z}^t$ and $\boldsymbol{q}_+^t$ (cf. Fig. 4b).

$$\mathcal{L}_t = \frac{1}{T}\sum_{t=1}^{T} -\log \frac{\exp(\boldsymbol{z}^t \cdot \boldsymbol{q}_+^t/\tau)}{\sum_{i=1}^{N}\sum_{s=1}^{S_{q_i}} \exp\left(\boldsymbol{z}^t \cdot \boldsymbol{q}_i^s/\tau\right)}. \tag{5}$$

This process enforces video features to learn the ability of temporally reasoning, thus leading to more localization-friendly video features.

Integrating the above constraints, our final loss function is as follows.

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_f \mathcal{L}_f + \lambda_t \mathcal{L}_t, \tag{6}$$

where λ_c, λ_f, and λ_t balance the focus on different constraints during training.

4 Experiments

4.1 Settings of Pre-training

Datasets. We pre-trained our model on three public datasets: **1)** HowTo100M [39]. It consists of more than 1.2M videos accompanied with ASR-generated speech transcription. The provided transcription is used to create video-sentence pairs separated by each timestamp. **2)** WebVid-2M [39]. It contains about 2.5M well-aligned web video-text pairs. **3)** Google Conceptual Captions [48]. It contains 3.3M image and description pairs harvested from the web.
Encoders. Following [4,54,67], we adopted ViT-B/16 [13] with space-time attention [5] as the video encoder. The spatial attention weights in the transformer were initialized with ImageNet-21k pre-trained weights while the temporal attention weights were set to zero. We chose a lightweight DistilBERT [47] as the language encoder. Following [4,41,52,3], the language encoder was initialized with the weights pre-trained on English Wikipedia and Toronto Book Corpus.
Implementation Details. For the video in each video-sentence pair, we sampled 8 clips of 16 frames equidistantly and fed them to the video encoder to obtain clip-level features. All frames were resized to 224×224. For downstream transfer, we extracted video features with the well-trained model in a dense manner, *i.e.*, every 16 consecutive frames were grouped to compute one clip feature.

Experiments were conducted on 64 V100 GPUs with a batch size of 256 and lasted for 200 epochs. We used Adam [32] with the initial learning rate 10^{-4} as the optimizer. The learning rate decayed by 0.1 at the 100^{th} and 160^{th} epoch. Random flip, random crop, and color jitter for video data augmentation were included. The loss balance factors λ_c, λ_f, and λ_t were set to 0.5, 1, 1, respectively. The temperature factor τ used in contrastive learning was set to 0.07 following [59,43] and K in Eq.(2) was set to 3. Features in all three contrastive losses were ℓ_2-normalized before computation.

4.2 Transfer Results on Video Retrieval

Datasets. We evaluate our LocVTP on the widely-used benchmark **MSR-VTT** dataset [64]. It is composed of 10K YouTube videos (9K for training and 1K for test). We report results on the train/test splits introduced in [69].
Results. 1) As can be seen, we achieve state-of-the-art performance under both sets of data, *i.e.*, HowTo100M and CC3M+WV2M. Specifically, when pre-trained on CC3M+WV2M, LocVTP outperforms Frozen [4] by an absolute lift of 4.8% on R@5. **2)** It should be pointed out that although using RGB data only, our LocVTP achieves better performance than the methods using multi-modal expert features including motion, face, and speech, *e.g.*, MMT [14]. **3)** The recent work CLIP [43] provides a stronger vision encoder and we also evaluate the performance based on it. It is shown that the CLIP's weights greatly improve the performance of LocVTP with R@5 achieving 72.8%, surpassing top-performing CLIP-based methods. **4)** Our LocVTP also outperforms previous methods under the zero-shot setting, showing its generalization ability.

Method	Vis Enc. Init.	Pre-trained Data	#pairs	R@1	R@5	R@10	MdR
UniVL [35]	-	HowTo100M	136M	21.2	49.6	63.1	6.0
ClipBERT [26]	-	COCO, VGen	5.6M	22.0	46.8	59.9	6.0
CE [31]	Multi-modal	HowTo100M	136M	20.9	48.8	62.4	6.0
MMT [14]	Multi-modal	HowTo100M	136M	26.6	57.1	69.6	4.0
HIT [30]	Multi-modal	HowTo100M	136M	30.7	60.9	73.2	2.6
Clip4clip† [36]	CLIP	HowTo100M	136M	44.5	71.4	81.6	2.0
VideoClip [63]	CLIP	HowTo100M	136M	30.9	55.4	66.8	-
OA-Trans [54]	CLIP	CC3M, WV2M	5.5M	40.9	70.4	80.3	2.0
Frozen [4]	ImageNet	CC3M, WV2M	5.5M	31.0	59.5	70.5	3.0
ActBERT [76]	VisGenome	HowTo100M	136M	16.3	42.8	56.9	10.0
SupportSet [41]	IG65M, ImageNet	HowTo100M	136M	30.1	58.5	69.3	3.0
HERO [27]	ImageNet, Kinetics	HowTo100M	136M	16.8	43.4	57.7	-
AVLnet [46]	ImageNet, Kinetics	HowTo100M	136M	27.1	55.6	66.6	4.0
NoiseEstimation [3]	ImageNet, Kinetics	HowTo100M	136M	17.4	41.6	53.6	8.0
DECEMBER [52]	ImageNet, Kinetics	HowTo100M	136M	30.7	60.9	73.2	2.6
OA-Trans [54]	ImageNet	CC3M, WV2M	5.5M	35.8	63.4	76.5	3.0
RegionLearner† [67]	ImageNet	CC3M, WV2M	5.5M	36.3	63.9	72.5	3.0
LocVTP (Ours)	**ImageNet**	**HowTo100M**	**136M**	**37.4**	**66.6**	**80.5**	**3.0**
LocVTP (Ours)	**CLIP**	**HowTo100M**	**136M**	**46.3**	**72.8**	**82.0**	**2.0**
LocVTP (Ours)	**ImageNet**	**CC3M,WV2M**	**5.5M**	**36.5**	**64.3**	**76.8**	**3.0**
		Zero-shot					
SupportSet [41]	IG65M, ImageNet	HowTo100M	136M	8.7	23.0	31.1	31.0
Frozen [4]	ImageNet	CC3M, WV2M	5.5M	18.7	39.5	51.6	10.0
OA-Trans [54]	ImageNet	CC3M, WV2M	5.5M	23.4	47.5	55.6	8.0
OA-Trans [54]	CLIP	CC3M, WV2M	5.5M	31.4	55.3	64.8	4.0
LocVTP (Ours)	**ImageNet**	**HowTo100M**	**136M**	**24.7**	**48.9**	**56.1**	**8.0**
LocVTP (Ours)	**CLIP**	**HowTo100M**	**136M**	**32.7**	**55.7**	**64.9**	**4.0**
LocVTP (Ours)	**ImageNet**	**CC3M,WV2M**	**5.5M**	**22.1**	**48.0**	**55.3**	**8.0**

Table 1: **Video retrieval performance on MSR-VTT.** Vis Enc. Init.: Datasets used for pre-training visual encoders. Methods using multi-modal features are grayed out. COCO: Coco Captions [10]; VGen: Visual genome [25]; CC3M: Conceptual captions [48]; WV2M: WebVid-2M [4]; † denotes the technical report available on ArXiv.

4.3 Transfer Results on Temporal Grounding

Settings. We validate the performance of pre-trained representations on temporal grounding, which aims to localize actions corresponding to the sentence from an untrimmed video. Specifically, we re-train the mainstream temporal grounding method 2D-TAN [74][6] by only replacing the original input features with pre-trained ones. For ease of feature extraction, we choose representative VTP methods with publicly-available codes for comparisons.

Datasets and Metrics. 1) ActivityNet Captions (ANet) [24]. It contains 20K untrimmed videos with 100K descriptions. By convention, we use 37,417 video-query pairs for training, 17,505 pairs for validation, and 17,031 pairs for testing. **2)** Charades-STA [15]. Following the official split, 12,408 video-query pairs are used for training, and 3,720 pairs for testing. **3)** TACoS [44]. It has 10,146 video-query pairs for training, 4,589 pairs for validation, and 4,083 pairs for testing.

[6] We choose 2D-TAN since it is relatively simple without too many dataset-specific parameters, which can fairly verify the effectiveness of pre-training features. Results on more advanced baselines are available in the supplementary material.

Models	PT Data	ANet Captions				Charades-STA				TACoS			
		$R_1^{0.5}$	$R_1^{0.7}$	$R_5^{0.5}$	$R_5^{0.7}$	$R_1^{0.5}$	$R_1^{0.7}$	$R_5^{0.5}$	$R_5^{0.7}$	$R_1^{0.3}$	$R_1^{0.5}$	$R_5^{0.3}$	$R_5^{0.5}$
Sep.Pre. [74]	Kinetics	44.4	27.1	77.6	62.1	39.7	23.8	79.6	52.3	37.2	25.6	58.2	45.5
LocVTP (Ours)	**HT‡**	**45.2**	**27.1**	**78.3**	**63.5**	**40.3**	**24.2**	**80.6**	**52.7**	**38.4**	**25.9**	**59.0**	**45.9**
VideoBERT* [49]	HT	37.2	21.0	66.7	53.6	32.7	19.5	68.1	46.2	33.8	22.2	51.6	41.0
MIL-NCE [38]	HT	41.8	24.5	73.5	57.7	37.0	21.2	74.3	50.4	35.1	23.5	53.7	42.5
UniVL [35]	HT	42.2	25.4	75.3	60.5	38.2	22.7	77.2	51.4	35.7	23.7	55.8	43.7
SupportSet* [41]	HT	41.9	25.2	74.7	58.3	37.4	21.6	75.6	50.9	35.5	23.5	54.2	43.2
LocVTP (Ours)	**HT**	**48.2**	**30.5**	**80.1**	**64.7**	**43.6**	**26.3**	**81.9**	**55.3**	**41.6**	**28.9**	**61.4**	**47.6**
Frozen [4]	CC,WV	43.3	25.8	75.8	59.3	38.8	22.9	77.6	50.3	35.7	23.5	54.4	43.7
OA-Trans* [54]	CC, WV	43.6	25.9	76.5	60.2	39.2	22.6	78.5	50.8	35.2	22.5	53.4	42.6
LocVTP (Ours)	**CC,WV**	**46.1**	**27.6**	**78.9**	**63.7**	**41.2**	**24.8**	**81.3**	**53.5**	**39.6**	**27.8**	**60.4**	**47.9**
December [52]	HT	43.0	25.1	76.0	60.2	37.2	21.6	78.3	50.6	34.8	22.9	55.1	43.9
ClipBERT [26]	CO,VG	42.6	24.6	75.3	59.7	37.0	20.8	77.7	50.2	33.7	21.0	54.3	43.3

Table 2: **Temporal grounding performances using pre-trained representations.** `Sep.Pre.`: separately pre-training, *i.e.*, the video encoder supervisedly pre-trained on Kinetics and text encoder taken from BERT. We retrain the the temporal grounding method 2D-TAN [74] using the pre-trained features. HT: HowTo100M; CO: Coco Captions [10]; VG: Visual genome [25]; CC: Conceptual captions [48]; WV: WebVid-2M [4]; **HT‡**: the subset of HowTo100M with the same training volume as Kinetics (300K pairs). Methods with * are not open source and we implement them ourselves. † denotes the technical report available on ArXiv.

Following prior works, we adopt "R@n, IoU@m" (abbreviated as R_n^m) as the metric, Specifically, R_n^m is defined as the percentage of at least one of top-n retrieved moments having IoU with the ground-truth moment larger than m.
Results. 1) As shown in Table 2, even trained with a much larger dataset, the current popular video-text pre-training frameworks achieve inferior performance compared to the separately pre-trained one. For example, Frozen [4] reaches 43.3% at $R_1^{0.5}$ on ANet Captions, which is 1.1% absolute value lower than the separately pre-trained counterpart. **2)** Either pre-trained on HowTo100M or CC + WV, our LocVTP outperforms both video-text pre-training methods by a large margin on all three datasets. For example, pre-trained on HowTo100M, LocVTP surpasses the separately pre-trained method by 3.8% on $R_1^{0.5}$ of ANet Captions. **3)** For more fair comparisons, we sample a subset of HowTo100M by selecting the same training sample as Kinetics [23] (300K training pairs), denoted as HT‡ in Table 2. Although using noisy ASR captions, the results demonstrates that under the same training data volume, our LocVTP still shows better performance compared to the separately pre-trained method. This manifests that our performance improvement is brought by the sound architecture design rather than just the use of the large-scale dataset.

4.4 Transfer Results on Action Step Localization

Settings. In action step localization, each video belongs to a task and is annotated with multiple action steps described with short natural languages. The goal is to align each frame with the correct step in the text form. Following [39,76,35,68], we take [77] as the downstream localization method. Specifically,

we compute the similarity between each frame and the action step descriptions in feature space to find the optimal frame-wise order of action steps for a video.

Datasets and Metrics. We experiment on the instructional video dataset CrossTask [77], which includes 83 tasks and 4.7K videos. Each task is described with an ordered list of steps with manual natural language descriptions. We perform the same evaluation protocol as in [77] by reporting the average recall (CTR).

Results. Table 3 reports the action step localization performance on CrossTask dataset. Our LocVTP pre-trained feature achieves state-of-the-art performance with CTR reaching 51.7%, surpassing the previous method VideoClip by 4.4%. Our competitive performance demonstrates that LocVTP features can effectively perceive detailed action steps.

Method	CTR	FA
Zhukov *et al.* [77]	31.6	-
NN-Viterbi [1]	–	21.2
CBT [38]	–	53.9
MIL-NCE [38]	40.5	61.0
ActBERT [76]	41.4	57.0
UniVL [35]	42.0	70.0
TACo [68]	42.5	68.4
VideoClip [63]	47.3	68.7
VLM [62]	46.5	68.4
LocVTP (Ours)	**51.7**	**72.9**

Table 3: Comparison results of action step localization (CTR: average recall) and action segmentation (FA: frame-wise accuracy).

4.5 Transfer Results on Action Segmentation

Settings. We assess our LocVTP on action segmentation, which aims to predict the action label frame-wisely for each video frame. It is a pure vision task without the use of the text encoder. Following [68,35,76], we encode the input video frames with the well-trained video encoder and apply a linear classifier upon the features to predict action labels.

Datasets and Metrics. We conduct experiments on the widely used COIN dataset [51] and the frame-wise accuracy (FA) is taken as the evaluation metric.

Results. As shown in Table 3, our LocVTP achieves state-of-the-art performance with FA reaching 72.9%. This further demonstrates the superiority of our feature in localization tasks even in the absence of language guidance.

4.6 Ablation Study on Training Objective[7]

Training Strategy. Coarse-grained contrastive alignment loss $\mathcal{L}_c$ provides a basic cross-modal matching prior and we introduce three potential ways to use it: **1)** *multi-stage training*: first perform coarse-grained training and then use the trained model to initialize other stages. **2)** *warm-up training*: decrease λ_c exponentially from 1 to 0 throughout the training process. **3)** *weighted training*: set λ_c to a constant value. Here we set $\lambda_c = 0.5$. As shown in Table 4a, we find the weighted training strategy achieves the best performance and warm-up training is slightly behind. Multi-stage training is the least effective one.

[7] If not specified, all ablation studies are conducted on the downstream temporal grounding task at ActivityNet Captions dataset. We use LocVTP pre-trained on HowTo100M with ImageNet initialization.

Mode	$R_1^{0.5}$	$R_1^{0.7}$
multi-stage	47.4	29.7
warm-up	47.7	30.1
weighted	**48.2**	**30.5**

(a)

	$\mathcal{L}_c$	$\mathcal{L}_f$	$\mathcal{L}_t$	$R_1^{0.5}$	$R_1^{0.7}$
#1	✓	✓	✓	**48.2**	**30.5**
#2	✓	✓		46.7$_{-1.5}$	29.4$_{-1.1}$
#3	✓		✓	46.8$_{-1.4}$	29.6$_{-0.9}$
#4	✓			45.6$_{-2.6}$	29.0$_{-1.5}$

(b)

Method	$R_1^{0.5}$	$R_1^{0.7}$
Sep.Pre. [29]	48.9	29.0
Frozen [4]	47.3	26.8
LocVTP	**53.9**	**34.6**

(c)

Table 4: **Ablations studies** of (a) training strategies; (b) loss component; (c) comparison results on temporal grounding method CSMGAN [29]. `Sep.Pre.`: separately pre-training, *i.e.*, the video encoder supervisedly pre-trained on Kinetics and text encoder taken from BERT.

Loss Component. We present the loss component ablations in Table 4b. As shown, both fine-grained loss $\mathcal{L}_f$ and temporal aware loss $\mathcal{L}_t$ are crucial. For example, compared to the full version (exp.#1), removing $\mathcal{L}_f$ and $\mathcal{L}_t$ brings about 1.4% and 1.5% performance degradation on the $R_1^{0.5}$ metric, respectively.
More downstream temporal grounding baselines. We take another temporal grounding method CSMGAN [29] as the downstream baseline. As shown in Table. 4c, our LocVTP pre-trained feature consistently benefits this more advanced baseline.

4.7 Ablations on Fine-grained Contrastive Loss[7]

Correspondence Discovery Strategies. We experiment four potential strategies to extract cross-modal correspondences: **1)** *random*: randomly select K words for each clip; **2)** *2d-topk*: select the most similar $K \times T$ clip-word pairs; **3)** *word-topk*: select the most similar K clips for each word; **4)** *clip-topk*: select the most similar K words for each clip, namely the method illustrated in Section 3.3. As indicated in Table. 5a, the *random* and *2d-topk* matching strategies are the two worst options. For the *word-topk* matching, it is also sub-optimal, which can be attributed to the possibility of introducing words without concrete meanings (*e.g.*, articles or pronouns) into matched pairs.
Number of Selected Pairs K. We further ablate the hyper-parameter K used in the *clip-topk* strategy. Table 5b shows that the performance saturates at $K = 3$ and slightly decreases for $K = 4$. We conjecture that this may be because too few words have vague meanings while too large K value leads to the inability to establish accurate correspondences.

4.8 Ablations on Temporal aware Contrastive Loss[7]

Context Projection Head Components. In Eq. (4), the warped feature is generated based on both the direction $\text{sgn}(\delta)$ and distance $|\delta|$. Here we investigate eliminating either of them to see the difference. We observe in Table. 5c that removing either component decreases the performance, which indicates that both the direction and distance of bias δ are crucial for feature warping.
Maximum Bias Distance δ_{max}**.** Here we ablate different values for δ_{max}. From Table 5d, we can see that $\delta_{max} = 4$ achieves the best performance. This may

	$R_1^{0.5}$	$R_1^{0.7}$
random	42.0	25.8
2d-topk	44.8	27.2
word-topk	47.0	28.7
clip-topk	**48.2**	**30.5**

(a)

K	$R_1^{0.5}$	$R_1^{0.7}$
1	46.3	28.8
2	47.5	29.7
3	**48.2**	**30.5**
4	48.0	29.8

(b)

sgn(δ)	$\lvert\delta\rvert$	$R_1^{0.5}$	$R_1^{0.7}$
✓	✓	**48.2**	**30.5**
✓	✗	47.3	29.3
✗	✓	47.1	29.0
✗	✗	46.2	28.1

(c)

δ_{max}	$R_1^{0.5}$	$R_1^{0.7}$
2	47.6	29.2
3	47.8	29.5
4	**48.2**	**30.5**
5	47.7	28.9

(d)

$\mathcal{L}_t$ Mode	$R_1^{0.5}$	$R_1^{0.7}$
intra-modal	47.7	29.8
cross-modal	**48.2**	**30.5**

(e)

Method	$Accu_o$	$Accu_d$
LocVTP (w/ $\mathcal{L}_t$)	**72.8**	**58.2**
LocVTP (w/o $\mathcal{L}_t$)	69.0	56.5
UniVL [35]	64.2	52.8
MIL-NCE [38]	61.3	51.4

(f)

Table 5: **Ablations studies** of (a) correspondence discovery strategies; (b) selected pair number K; (c) context projection head. sgn(δ), $|\delta|$ denotes the direction and distance; (d) the maximum bias distance; (e) intra-modal *v.s.* cross-modal $\mathcal{L}_t$; (f) linear localization accuracy. $Accu_o$ and $Accu_d$ are order and distance prediction accuracy.

be because that small bias makes the model unable to perceive enough context, while a large bias makes contextual reasoning too difficult.

Intra-modal *v.s.* Cross-modal Constraint. In Section. 3.4, given the matched clip-word pair $\{\boldsymbol{v}^t, \boldsymbol{q}_+^t\}$ and the warped feature $\boldsymbol{z}^t$, we force the *cross-modal* supervision, *i.e.*, $\boldsymbol{z}^t \leftrightarrow \boldsymbol{q}_+^t$. Here, we apply the temporal aware contrastive loss $\mathcal{L}_t$ in a *intra-modal* manner which regards $\boldsymbol{z}^t$ and $\boldsymbol{v}^t$ as positive pairs, *i.e.*, $\boldsymbol{z}^t \leftrightarrow \boldsymbol{v}^t$. The results in Table 5e show that our adopted cross-modal mode outperforms the intra-modal one.

Temporal Sensitivity Analysis. As a sanity check, we devise two proxy tasks to evaluate the temporal sensitivity of pre-trained video features. As shown in Fig. 5a, n equidistantly sampled clips from one video are fed into the frozen video backbone to extract their corresponding features. Two linear classifiers are trained to perform two tasks: *order prediction* and *distance estimation*. The first task predicts the temporal index while the second one estimates the temporal distance of two clips. The results in Table 5f show that our LocVTP with temporal aware loss $\mathcal{L}_t$ outperforms the variant without it as well as two typical VTP methods (*i.e.*, UniVL and MIL-NCE), which shows that $\mathcal{L}_t$ clearly contributes to the localization ability.

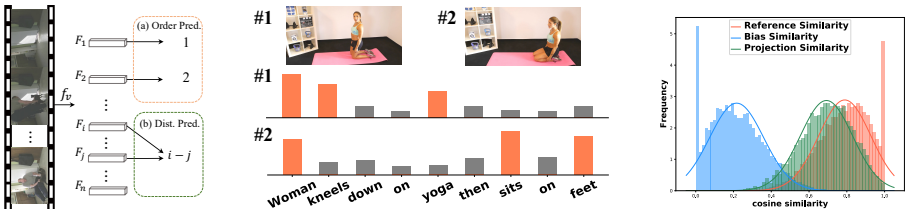
4.9 Visualization[8]

Cross-modal Correspondence Visualizations. Fig. 5b shows two frames[9] and their corresponding similarity scores with caption words. The top K highest scored words are marked with red ($K = 3$). Frame #1 and frame #2 have similar appearance views yet correspond to different action processes. Our method pinpoints the subtle differences and accurately finds the most relevant words.

UMAP Visualizations. As shown in Fig. 6, we provide UMAP [37] visualizations for *fused* multi-modal features, which are generated by multiplying the

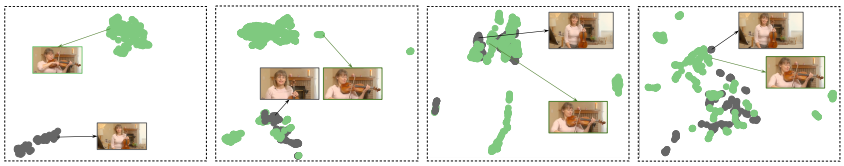
[8] More visualizations are left in the supplementary materials.

[9] Here we use "frame" to indicate the center frame of a video snippet.



(a) (b) (c)

Fig. 5: (a) Linear localization evaluations including order and distance prediction; (b) Cross-modal correspondence visualizations. Top K responsive words are marked with **red**. (c) Gaussian distributions of the **reference**, **biased**, and **projected** similarities.



(a) Ours($w/\mathcal{L}_t$) (b) Ours($w/o\mathcal{L}_t$) (c) UniVL (d) MIL-NCE

Fig. 6: UMAP visualizations. Clips corresponding to ground-truth caption are marked with **green** while others are with **gray**.

extracted video feature by one query feature. With the temporal aware loss $\mathcal{L}_t$, our LocVTP shows more separable distributions compared with LocVTP w/o $\mathcal{L}_t$, manifesting that $\mathcal{L}_t$ helps distinguish action-of-interest from background.

Similarity Distribution Visualizations. In Eq.(4), context projection head warps contextual clip $\boldsymbol{v}^{t+\delta}$ to the reference one $\boldsymbol{v}^t$. Here we collect 10K paired training samples and compute three sets of cosine similarities: reference similarity $(\boldsymbol{v}^t, \boldsymbol{q}_+^t)$, bias similarity $(\boldsymbol{v}^{t+\delta}, \boldsymbol{q}_+^t)$, and projection similarity $(\boldsymbol{z}^t, \boldsymbol{q}_+^t)$. Fig. 5c plots the histogram of these similarities. We can see that the distribution of projection similarity is close to that of reference similarity while far away from that of bias similarity. This demonstrates that our context projection head can effectively warp contextual features conditioned on the temporal information.

5 Conclusions

In this paper, we propose LocVTP, the first video-text pre-training framework for temporal localization tasks. Specifically, we apply cross-modal contrastive learning at both coarse-grained video-sentence and fine-grained clip-word levels. Besides, we propose a context warping pretext task and a temporal aware contrastive loss to enhance the temporal awareness of video features. Experimental results show that LocVTP achieves state-of-the-art performance when transferred to both retrieval-based and localization-based downstream tasks.

Acknowledgements. This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No: GXWD20201231165807007-20200814115301001).

References

1. Alayrac, J.B., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., Lacoste-Julien, S.: Unsupervised learning from narrated instruction videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4575–4583 (2016) 11
2. Alwassel, H., Giancola, S., Ghanem, B.: Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3173–3183 (2021) 4
3. Amrani, E., Ben Ari, R., Rotman, D., Bronstein, A.: Noise estimation using density estimation for self-supervised multimodal learning. arXiv preprint arXiv:2003.03186 (2020) 8, 9
4. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. arXiv preprint arXiv:2104.00650 (2021) 1, 4, 8, 9, 10, 12
5. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? arXiv (2021) 8
6. Cao, M., Chen, L., Shou, M.Z., Zhang, C., Zou, Y.: On pursuit of designing multi-modal transformer for video grounding. EMNLP (2021) 6
7. Cao, M., Zhang, C., Chen, L., Shou, M.Z., Zou, Y.: Deep motion prior for weakly-supervised temporal action localization. arXiv preprint arXiv:2108.05607 (2021) 6
8. Chen, L., Lu, C., Tang, S., Xiao, J., Zhang, D., Tan, C., Li, X.: Rethinking the bottom-up framework for query-based video localization. In: AAAI. pp. 10551–10558 (2020) 6
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020) 4
10. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 9, 10
11. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV. pp. 104–120 (2020) 4
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 2
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 4, 8
14. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 214–229. Springer (2020) 8, 9
15. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: ICCV. pp. 5267–5275 (2017) 2, 4, 9
16. Ging, S., Zolfaghari, M., Pirsiavash, H., Brox, T.: Coot: Cooperative hierarchical transformer for video-text representation learning. Advances in neural information processing systems **33**, 22605–22618 (2020) 2
17. Han, N., Chen, J., Xiao, G., Zhang, H., Zeng, Y., Chen, H.: Fine-grained cross-modal alignment network for text-video retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3826–3834 (2021) 4

18. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020) 4, 5, 6
19. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition. pp. 84–92. Springer (2015) 4
20. Hu, R., Singh, A.: Transformer is all you need: Multimodal multitask learning with a unified transformer. arXiv (2021) 4
21. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2758–2766 (2017) 2, 4
22. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014) 1
23. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv (2017) 1, 2, 10
24. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: ICCV. pp. 706–715 (2017) 2, 9
25. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV pp. 32–73 (2017) 9, 10
26. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7331–7341 (2021) 1, 4, 9, 10
27. Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. arXiv preprint arXiv:2005.00200 (2020) 9
28. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv (2019) 4
29. Liu, D., Qu, X., Liu, X.Y., Dong, J., Zhou, P., Xu, Z.: Jointly cross-and self-modal graph attention network for query-based moment localization. In: ACM MM. pp. 4070–4078 (2020) 12
30. Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., Wang, Z.: Hit: Hierarchical transformer with momentum contrast for video-text retrieval. arXiv preprint arXiv:2103.15049 (2021) 1, 4, 9
31. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487 (2019) 1, 4, 9
32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv (2017) 8
33. Lu, C., Chen, L., Tan, C., Li, X., Xiao, J.: Debug: A dense bottom-up grounding approach for natural language video localization. In: EMNLP. pp. 5147–5156 (2019) 6
34. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019) 4
35. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020) 9, 10, 11, 13

36. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021) 9
37. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018) 13
38. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889 (2020) 2, 10, 11, 13
39. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2630–2640 (2019) 1, 2, 8, 10
40. Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W.: Videomoco: Contrastive video representation learning with temporally adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11205–11214 (2021) 6
41. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. arXiv preprint arXiv:2010.02824 (2020) 4, 8, 9, 10
42. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2021) 6
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) 8
44. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. TACL pp. 25–36 (2013) 9
45. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3202–3212 (2015) 2, 4
46. Rouditchenko, A., Boggust, A., Harwath, D., Chen, B., Joshi, D., Thomas, S., Audhkhasi, K., Kuehne, H., Panda, R., Feris, R., et al.: Avlnet: Learning audio-visual language representations from instructional videos. arXiv preprint arXiv:2006.09199 (2020) 9
47. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019) 8
48. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL. pp. 2556–2565 (2018) 8, 9, 10
49. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7464–7473 (2019) 1, 10
50. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: EMNLP (2019) 4
51. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: Coin: A large-scale dataset for comprehensive instructional video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1207–1216 (2019) 2, 4, 11

52. Tang, Z., Lei, J., Bansal, M.: Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2415–2426 (2021) 8, 9, 10
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 4
54. Wang, A.J., Ge, Y., Cai, G., Yan, R., Lin, X., Shan, Y., Qie, X., Shou, M.Z.: Object-aware video-language pre-training for retrieval. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 1, 8, 9, 10
55. Wang, W., Zhang, M., Chen, R., Cai, G., Zhou, P., Peng, P., Guo, X., Wu, J., Sun, X.: Dig into multi-modal cues for video retrieval with hierarchical alignment. IJCAI (2021) 4
56. Wang, X., Zhu, L., Yang, Y.: T2vlad: global-local sequence alignment for text-video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5079–5088 (2021) 4
57. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3024–3033 (2021) 6
58. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2840–2848 (2017) 4
59. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018) 8
60. Xiao, S., Chen, L., Shao, J., Yueting, Z., Xiao, J.: Natural language video localization with learnable moment proposals. In: EMNLP (2021) 6
61. Xiao, S., Chen, L., Zhang, S., Ji, W., Shao, J., Ye, L., Xiao, J.: Boundary proposal network for two-stage natural language video localization. In: AAAI (2021) 6
62. Xu, H., Ghosh, G., Huang, P.Y., Arora, P., Aminzadeh, M., Feichtenhofer, C., Metze, F., Zettlemoyer, L.: Vlm: Task-agnostic video-language model pre-training for video understanding. arXiv preprint arXiv:2105.09996 (2021) 11
63. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084 (2021) 9, 11
64. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016) 2, 4, 8
65. Xu, M., Pérez-Rúa, J.M., Escorcia, V., Martinez, B., Zhu, X., Zhang, L., Ghanem, B., Xiang, T.: Boundary-sensitive pre-training for temporal localization in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7220–7230 (2021) 4
66. Xu, M., Perez Rua, J.M., Zhu, X., Ghanem, B., Martinez, B.: Low-fidelity video encoder optimization for temporal action localization. Advances in Neural Information Processing Systems **34** (2021) 4
67. Yan, R., Shou, M.Z., Ge, Y., Wang, A.J., Lin, X., Cai, G., Tang, J.: Video-text pre-training with learned regions. arXiv preprint arXiv:2112.01194 (2021) 1, 8, 9
68. Yang, J., Bisk, Y., Gao, J.: Taco: Token-aware cascade contrastive learning for video-text alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11562–11572 (2021) 10, 11

69. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 471–487 (2018) 8
70. Yuan, Y., Lan, X., Wang, X., Chen, L., Wang, Z., Zhu, W.: A closer look at temporal sentence grounding in videos: Datasets and metrics. arXiv (2021) 6
71. Zhang, C., Cao, M., Yang, D., Chen, J., Zou, Y.: Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In: CVPR. pp. 16010–16019 (2021) 6
72. Zhang, C., Cao, M., Yang, D., Jiang, J., Zou, Y.: Synergic learning for noise-insensitive webly-supervised temporal action localization. Image and Vision Computing **113**, 104247 (2021) 6
73. Zhang, C., Yang, T., Weng, J., Cao, M., Wang, J., Zou, Y.: Unsupervised pretraining for temporal action localization tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14031–14041 (2022) 4
74. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: AAAI. pp. 12870–12877 (2020) 2, 9, 10
75. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018) 2
76. Zhu, L., Yang, Y.: Actbert: Learning global-local video-text representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8746–8755 (2020) 9, 10, 11
77. Zhukov, D., Alayrac, J.B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3537–3545 (2019) 2, 4, 10, 11