

# SRL-CLIP: Efficient CLIP Video Adaptation via Structured Semantic Role Labels

Darshan Singh<sup>1</sup>, Zeeshan Khan<sup>2</sup>, Makarand Tapaswi<sup>1</sup>

<sup>1</sup>CVIT, IIT Hyderabad    <sup>2</sup>Inria, École normale supérieure, CNRS, PSL Research University

## Abstract

Adapting CLIP for videos has gained popularity due to its semantic and rich representation. While CLIP is a good starting point, it typically undergoes post-pretraining (contrastive finetuning) on large video narration or caption datasets (e.g. *HowTo100M*, *WebVid2.5M*). However, such narrations or captions often lack comprehensive information needed to represent a video holistically. As the learning signal from text is sparse, the visual learning is inefficient and adaptation requires millions of samples to post-pretrain.

In this work, we ask: is it possible to efficiently adapt CLIP for general and holistic video understanding? We use videos labeled with structured and dense Semantic Role Labels (SRLs) that capture actions, people or objects, their attributes, adverbs (manner), and location in a structured format representing the entire video in a holistic way. We generate rule-based captions from SRLs and demonstrate that simple contrastive finetuning on a **mere 23k** video-caption pairs is adequate to learn powerful, transferable representations applicable across a diverse range of video understanding tasks that require varying levels of perceptual granularity. Our adapted CLIP model, *SRL-CLIP*, exhibits comparable or superior performance on zero-shot text-to-video retrieval compared to state-of-the-art models that possess 4–8× more parameters and are post-pretrained on up to 6000× more data. *SRL-CLIP* surpasses CLIP on multiple video benchmarks, underscoring the efficient learning and improved representations.

## 1. Introduction

Large-scale vision language pretraining has proved effective in learning generalized and transferable visual representations, with powerful zero-shot capabilities [26–28, 44]. Among them, Contrastive Language-Image Pretraining (CLIP) [44] is a popular approach for learning rich semantic representations, thanks to the image-text alignment. As the representations generalize well, CLIP sees wide adoption in several downstream tasks ranging from simple classification to being used in Vision-Language Models

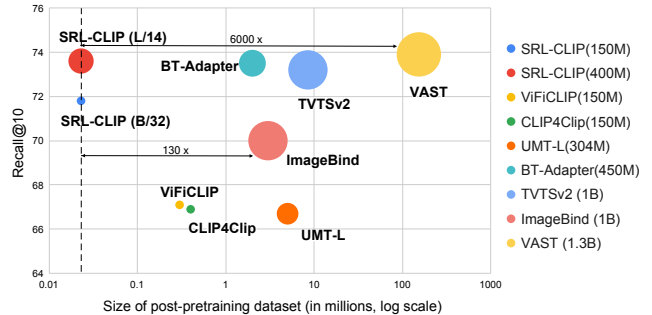


Figure 1. Zero-shot text-to-video retrieval performance on the MSRVT dataset. We compare SRL-CLIP (Ours) against various CLIP-based approaches that use orders of magnitude more post-pretraining data samples and/or have significantly larger models.

(VLMs) [1, 54, 56, 60]. This has spurred a lot of interest in adapting CLIP for video understanding [30, 33, 36, 45]. While these methods achieve promising results, they often come at the expense of large scale post-pretraining on hundreds of millions of videos.

We argue that as CLIP is pretrained on 400+ million image-language pairs, it already has the world knowledge required for general visual perception, and post-pretraining on hundreds of millions of videos can be wasteful. We hypothesize that the need for such large-scale post-pretraining is due to the contradiction between rich videos and sparse descriptions. Videos are a highly complex modality with multidimensional information about people, objects, states, actions, relations, that change over time. But, the textual description representing the video is generally very sparse. Narrations [35] or captions [3] often fail to capture details and therefore the capacity for visual processing is underutilized. The adaptation process requires millions of samples as it only extracts sparse information from each.

In this work, we revisit the large-scale post-pretraining paradigm for CLIP video adaptation and propose to use small-scale, but dense captions for efficient and holistic representation learning. Specifically, we use the VidSitu dataset [47] that is annotated using Semantic Role Labels (SRLs) that capture the holistic situation. In VidSitu, every 10 second video clip is divided into 2 second *events*. Each event contains structured annotations answering who (subject) is doing what (action), to/with whom (patient), where

(scene), how (manner, adverbs), and why (purpose, goal). Such details in a video description represent the visual concepts well and can provide a strong learning signal.

We propose to adapt CLIP through alignment between video-text SRLs, highlighting the power of rule-based high quality prompts generated from SRL. Our approach has multiple advantages: (i) The structured nature of SRL-based prompts (who, what, where, *etc.*) facilitate learning holistic visual concepts consistently across all the videos. (ii) Vid-Situ videos have many shot changes, while the annotated captions may represent visual concepts across all the events. This presents a challenging scenario with a strong signal to learn object permanence and temporal consistency, by meaningfully combining information across frames and events. (iii) Multiple events from the same videos show small differences creating *natural hard negatives* for learning. In addition, due to the structured nature of SRL, we can add/swap/replace verbs, nouns, or roles in an SRL-based prompt, easily creating *hard negative text prompts* that further boost the model’s performance.

We show that SRL-CLIP, adapted on a small dataset with dense annotations, learns powerful general purpose representations, improving over base CLIP on multiple video-language tasks. In particular, we achieve these improvements through *efficient adaptation* – a ViT-B/32 with 150M params, post-pretraining dataset of 23k video clips, trained on a single 12GB RTX 2080 GPU, for only 5 hours!

**Contributions.** In summary, we: (i) Favor efficient post-pretraining using small-scale but dense prompts, as compared to the current inefficient trend of using large-scale datasets. (ii) Propose to use SRL to create rule-based dense captions that capture holistic concepts in a video, through details such as actions, persons, objects, attributes, relations, manner, location, reasons, *etc.* (iii) Show that post-pretraining CLIP on mere 23k videos with dense SRL captions is efficient and achieves on-par or better performance on zero-shot text-to-video retrieval (see Fig. 1), compared to CLIP based video models that have 4–8× more parameters and are trained on upto 6000× more data. (iv) Outperform the original CLIP model on various tasks: video situation recognition, dense video captioning and localization, and text-to-video retrieval, demonstrating that representations learned with dense holistic captions generalize well across multiple tasks requiring different perceptual granularity.

We emphasize that our innovation lies in an efficient and effective recipe to adapt CLIP for video data rather than architectural or modeling modifications.

### 1.1. Small Data Statement

Our work qualifies as small data research because the entire post-pretraining of SRL-CLIP is performed on only 23k video clips from VidSitu [47]. This is in stark contrast to existing CLIP video adaptation methods that rely

on datasets that are orders of magnitude larger than ours as shown in Tab. 1. Our training is also carried out on a single 12GB RTX 2080 GPU in just 5 hours for the ViT-B/32 model, achieving competitive or superior results to models with up to 4 – 8× more parameters and trained on orders of magnitude more data.

To tackle the small data challenge, we employ the following methods: (i) *Dense structured annotations*: We leverage Semantic Role Labels (SRLs) that capture who, what, whom, where, how, and why for each event, providing a much richer per-sample learning signal than sparse narrations or captions. This compensates for the small dataset size by maximizing the information extracted from each video. (ii) *Parameter-efficient adaptation*: LoRA modules [19] are added to a frozen CLIP backbone, preventing catastrophic forgetting while enabling effective adaptation with limited data. (iii) *Hard negatives from structure*: The compositional nature of SRL allows us to systematically create hard negatives by swapping verbs, nouns, or roles, boosting fine-grained understanding without requiring additional data. (iv) *Video contextualizer*: A lightweight temporal module contextualizes frame-level features across events, enabling fine-grained temporal understanding from limited videos.

## 2. Related Work

Several attempts have been made to adapt CLIP for videos [5, 10, 33, 34, 42, 45, 53, 55, 58, 62]. However, they are primarily focused on task-specific adaptation for action recognition [45, 53] or text-to-video retrieval [10, 33, 34, 55, 58, 62]. Task specific adaptations may lead to the loss of generalized representations. Different from above, we propose task-agnostic adaptation of CLIP, and focus on extending CLIP to videos using holistic video understanding datasets. This allows our model to be applied on a diverse range of video understanding tasks that demand different levels of perceptual granularity. We discuss related works in two dimensions: (i) datasets used for adapting CLIP; and (ii) approaches for adapting CLIP for videos.

**1. Popular VL datasets for adapting CLIP.** The unparalleled success of training Large Language Models (LLMs), *e.g.* GPT2 [43], LLaMA [50], with massive datasets has brought similar trends to the vision community.

**Towards text-to-video retrieval.** Large-scale web scraped datasets such as HowTo100M [35] and HD-VILA [61] align video clips with narrations. More recently, WebVid-2.5M [3] was curated from stock footage with textual descriptions resulting in better captions, that align with the video. These datasets are used by several methods for adapting CLIP or training a video-text contrastive alignment approach from scratch [12, 33, 40, 58, 62].

**Towards action understanding.** Datasets typically consist of 10-second videos annotated with a single action. To adapt

CLIP using supervised video-text contrastive learning, the widely used idea is to create CLIP-like prompts from the action labels [34, 45, 53] on datasets like Kinetics [7] or Something-Something [16].

**Towards holistic video understanding.** Action labels [7] or narrations [35] fail to capture the complex, hierarchical, multifaceted aspects of videos. To learn holistic and fine-grained representations, we propose to exploit the SRL annotations in VidSitu [47]. We observe that SRL-CLIP consistently outperforms base CLIP on multiple video understanding tasks that require varying degrees of granularity, indicating that it has improved holistic reasoning abilities. Moreover, a model trained on SRL captions outperforms methods that are post-pretrained on large-scale web video datasets (often 2-3 orders of magnitude larger) [3, 35, 61] on *zero-shot* video retrieval, indicating improved vision-language alignment.

**2. Approaches for adapting CLIP** for videos can be classified into direct fine-tuning or through adapters.

**Fine-tuning approaches** typically follow frame-level feature extraction from CLIP followed by temporal aggregation. The resulting video feature is aligned with the corresponding text prompt via contrastive loss. Either partial, or all the parameters of CLIP are fine-tuned [45, 53]. While a majority of the methods [10, 33, 34, 53] use a Transformer [51] for temporal aggregation of frame-level features, simple mean pooling has been shown to be effective for a narrow task of action recognition [45]. Other approaches use a weighted-mean of frame embeddings based on query-scoring [4]; compute frame-level attention based on text [15]; integrate temporal aggregation within the image encoder [62]; or suggest using a temporal model in parallel to the image encoder [41].

**Adapters**, the alternative to finetuning, are lightweight modules injected between layers of a pretrained model for efficient adaptation on a downstream task [18, 19, 39, 49]. The original parameters are usually frozen, and only the adapter is trained, allowing for efficient adaptation. Prior works in CLIP adaptation have used spatial adapters [13], spatio-temporal adapters [38], and cross-modal adapters [21] for efficient adaptation to downstream tasks. Recently there has been a surge of methods using low-rank adapters (LoRA) [19] and advances [17, 31] for their high efficiency enabled by low rank learnable matrices during training and no computational overhead during inference. We follow this approach instead of fine-tuning, allowing us to efficiently post-pretrain CLIP’s ViT-B/32 image and/or text encoders on a single 12GB GPU.

### 3. Adaptation: From CLIP to SRL-CLIP

We present our approach to post-pretrain the CLIP model on VidSitu, a densely annotated video dataset. We start with background information related to CLIP (Sec. 3.1), followed by our adaptation strategy with specific emphasis on the

architecture modifications for training and SRL prompts (Sec. 3.2). We end this section with a discussion on how SRL prompts support creation of difficult negatives (Sec. 3.3) and some details about the adaptation process (Sec. 3.5).

#### 3.1. Preliminaries

Consider a batch  $B$  of paired image-text data:  $\{(f_i, t_i)\}_{i=1}^B$ , where  $f_i$  is the image and  $t_i$  describes  $f_i$ . The CLIP model [44] consists of an image encoder  $\mathbf{f}_i = \Phi_I(f_i)$  and a text encoder  $\mathbf{t}_i = \Phi_T(t_i)$  that are trained in a contrastive manner by applying the InfoNCE loss [37]:

$$L(\mathbf{f}_i, \mathbf{t}_i) = -\log \frac{\exp(\mathbf{f}_i^T \mathbf{t}_i)}{\sum_{j=1}^B \exp(\mathbf{f}_i^T \mathbf{t}_j)}, \quad (1)$$

and the corresponding symmetric version  $L(\mathbf{t}_i, \mathbf{f}_i)$ . The loss for the entire batch is  $\mathcal{L} = \sum_{i=1}^B (L(\mathbf{f}_i, \mathbf{t}_i) + L(\mathbf{t}_i, \mathbf{f}_i))$ .

Previous works have adapted CLIP using video-text pairs by mean pooling across multiple video frames [45]. Instead of an image  $f_i$ , consider a video-text pair  $(V_i, t_i)$  where  $V_i = \{f_{ij}\}_{j=1}^{L_i}$  has  $L_i$  frames. We can adapt CLIP by computing a video representation  $\mathbf{v}_i = \text{mean}_j(\mathbf{f}_{ij})$  and using the same loss  $L(\mathbf{v}_i, \mathbf{t}_i)$ .

**Video contextualizer (VC).** Instead of mean pooling, a VC may be used for learning video-text representations [33, 34]. To contextualize all frames, a simple Transformer encoder (Tx) [51] ingests frame representations as tokens. A learnable CLS token,  $\mathbf{h}_{\text{CLS}}$ , is inserted before all frames. The input to the VC is:  $[\mathbf{h}_{\text{CLS}}, \mathbf{f}_{i1}, \dots, \mathbf{f}_{iL_i}]$ . Position encoding [51] is added to the video frame tokens to specify their temporal order. Finally, the output at the CLS token is considered as the video representation, *i.e.*,  $\mathbf{v}_i = \hat{\mathbf{h}}_{\text{CLS}}$ . Note, the VC can be trained jointly with adaptation of the backbone through the same loss  $L(\mathbf{v}_i, \mathbf{t}_i)$ .

Next, we show how VC can be used with VidSitu.

#### 3.2. Adaptation with SRLs

The VidSitu dataset [47] features videos that are split into  $P=5$  contiguous short *events*, *i.e.*  $V_i = [E_{ik}]_{k=1}^P$ . Each event contains a detailed annotation: an action label and corresponding semantic role labels (SRL) with the role-noun pairs. For example, in Fig. 3, we show the action *drive*, with roles *driver*, *vehicle*, *manner* (of driving), and *scene*, each described through a short caption (noun).

We use such fine-grained labels to create a prompt for each event,  $t_{ik}$ , leading to event and text pairs  $(E_{ik}, t_{ik})$ . Fig. 2 illustrates the overall adaptation strategy.

**Video contextualizer (VC) for encoding VidSitu events.** We modify the VC to account for VidSitu’s structured annotations. During training,  $B$  videos  $\{V_i\}_{i=1}^B$  are fed to the model at once. Each video is split into  $P$  events. From each event, we sub-sample  $T$  frames, *i.e.*, for each video, we have  $L=P \cdot T$  frames. The VC operates over a sequence of all

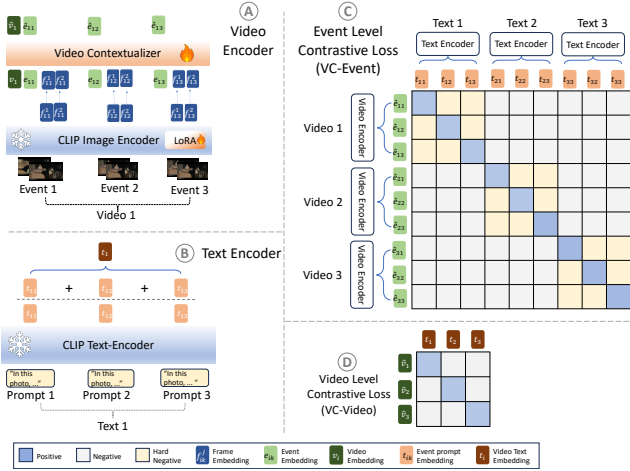


Figure 2. Overview of our CLIP adaptation strategy. **Top-left (A)** shows the visual encoder, consisting of the CLIP backbone and the video contextualizer (VC), applied to a single video with  $P=3$  events for illustration. **Bottom-left (B)** shows the frozen CLIP text encoder extracting event-level representations. **Top-right (C)** shows the event-level contrastive loss, VC-Event, with natural hard negatives due to multiple events within a single video. **Bottom-right (D)** shows the video-level contrastive loss, VC-Video.

frames, passed through the CLIP image encoder. Let  $f_{ik}^j$  be the  $j^{\text{th}}$  frame for event  $E_{ik}$  of video  $V_i$ .  $\mathbf{v}_i$  represents the overall video  $V_i$  and  $\mathbf{e}_{ik}$  represents event  $E_{ik}$ .

Similar to CLS in BERT [9], we create two learnable token types that collect video information. Note, these embeddings are shared across all videos. To indicate the type of token, we augment visual/learnable encodings with type embeddings  $\mathbf{e}_v^{\text{typ}}$  for video,  $\mathbf{e}_e^{\text{typ}}$  for event, and  $\mathbf{e}_f^{\text{typ}}$  for the frame. Furthermore, we encode position with two embeddings,  $\mathbf{e}^{\text{e-pos}}$  for event position and  $\mathbf{e}^{\text{f-pos}}$  for frame position within the event. Overall, our input tokens are:

$$\mathbf{v}_i = \mathbf{v}_i + \mathbf{e}_v^{\text{typ}}, \quad (2)$$

$$\mathbf{e}_{ik} = \mathbf{e}_{ik} + \mathbf{e}_e^{\text{typ}} + \mathbf{e}_k^{\text{e-pos}}, \quad (3)$$

$$\mathbf{f}_{ik}^j = \mathbf{f}_{ik}^j + \mathbf{e}_f^{\text{typ}} + \mathbf{e}_k^{\text{e-pos}} + \mathbf{e}_j^{\text{f-pos}}, \quad (4)$$

and passed to the VC,  $\Phi_V$ , after LayerNorm [2]:

$$\Phi_V([\mathbf{v}_i, \mathbf{e}_{i1}, \mathbf{f}_{i1}^1, \dots, \mathbf{f}_{i1}^T, \dots, \mathbf{e}_{iP}, \mathbf{f}_{iP}^1, \dots, \mathbf{f}_{iP}^T]). \quad (5)$$

We denote outputs after the VC as  $\hat{\mathbf{v}}_i, \hat{\mathbf{e}}_{ik}, \hat{\mathbf{f}}_{ik}^j$  for video, event, and frame tokens respectively. Fig. 2 (top-left) illustrates this process.

**Creating event-level prompts.** Our prompt  $t_{ik}$  is a simple template that enumerates over the action and semantic role labels for each event. An example is shown below. Template words in gray, label type in *italics*, the label in line:

In this photo, the *action* is walk where, the *walker* is man with short hair wearing collared shirt, *di-*

*rection* is forward, *manner* is slowly, and *scene* of the event is apartment.

We also consider generating natural language prompts using a language model (LLaMa [50]). Labels are underlined:

In this photo, a man with short hair wearing a collared shirt is walking slowly in an apartment.

However, these show worse performance. Telling the model that walking is the *action*, or the role played by a person with collared shirt is *walker*, and the apartment is a *scene* allows for holistic and dense representation learning.

**Losses.** We train our model with losses at multiple levels. For this part, we will recall some notations: the prompt encoding  $\mathbf{t}_{ik} = \Phi_T(t_{ik})$ ;  $\mathbf{f}_{ik}^j$  is the frame encoding before VC;  $\hat{\mathbf{e}}_{ik}$  is the event encoding after VC; and  $\hat{\mathbf{v}}_i$  the video encoding after VC. We also consider a prompt representation for the full video obtained by mean pooling over all event-level prompts,  $\mathbf{t}_i = \text{mean}_k(\mathbf{t}_{ik})$ .

All losses below follow the contrastive loss (see Eq. (1)):

(i) **CLIP-Event** applies a loss on the event representation obtained by mean pooling raw CLIP frame encodings:  $L_{\text{event}}^{\text{CLIP}} = L(\text{mean}_j(\mathbf{f}_{ik}^j), \mathbf{t}_{ik})$ .

(ii) **CLIP-Video** applies a loss on the video representation obtained by mean pooling raw CLIP frame encodings across the video and contrasting against the video-level prompt:  $L_{\text{video}}^{\text{CLIP}} = L(\text{mean}_{jk}(\mathbf{f}_{ik}^j), \mathbf{t}_i)$ .

(iii) **VC-Event** applies a loss on the event representation post VC and the event-level prompt:  $L_{\text{event}}^{\text{VC}} = L(\hat{\mathbf{e}}_{ik}, \mathbf{t}_{ik})$ . Fig. 2 (top-right) represents this loss. Note how multiple events from the same video are used as negatives.

(iv) **VC-Video**: applies a loss on the video representation post VC and video-level prompt:  $L_{\text{video}}^{\text{VC}} = L(\hat{\mathbf{v}}_i, \mathbf{t}_i)$ . Fig. 2 (bottom-right) represents this loss.

We also use the symmetric version of the losses, e.g.  $L(\mathbf{t}_{ik}, \hat{\mathbf{e}}_{ik})$ , which are not shown here for brevity. We train the VC and adapt the backbone through a combination of all losses:  $\mathcal{L} = L_{\text{event}}^{\text{CLIP}} + L_{\text{event}}^{\text{VC}} + \lambda(L_{\text{video}}^{\text{CLIP}} + L_{\text{video}}^{\text{VC}})$ .

### 3.3. SRL Contains and Facilitates Hard Negatives

Contrastive learning requires hard negatives (HN) during training to prevent the model from finding the easy difference between the image and negative prompt.

**Natural hard negatives.** Training with dense template prompts along with our batch creation strategy gives us natural hard negatives which promotes fine-grained, holistic, and efficient representation learning. For the CLIP-Event and VC-Event losses, negative prompts are obtained from different events of the same video which are generally quite similar (perhaps differing in only some of the verb/roles/nouns). Fig. 2 top-right shows these natural HNs in light yellow background. For example, a template prompt for a different event of the same video as described above (walking example):

In this photo, the *action* is sit where, the *thing sitting* is man with short hair wearing collared shirt, *manner* is casually, and *scene* of the event is apartment.

The subtle differences in verbs, roles, or their corresponding nouns, presents a challenging learning scenario for the model, resulting in multi-granularity representations.

### Artificial hard negatives by replacing verb-role pairs.

Although natural HNs provide challenging samples to the model, template-based prompts can be used easily to create artificial HNs by replacing verb-role pairs. Starting from the positive prompt, we replace the verb with a randomly sampled verb from the batch. We also replace its corresponding roles, but keep the nouns unchanged. Note, common roles such as *direction*, *manner*, *scene* remain unchanged; making the prompts quite hard. This allows the model to focus more on the *action*. We use  $\mathcal{N}_{vr}$  such negatives. Differences to our walking example are in red:

In this photo, the *action* is jog where, the *jogger* is man with short hair wearing collared shirt, *direction* is forward, and *scene* of the event is apartment.

**Incorporating negatives in the loss.** HNs are only added to the event-level losses (VC-Event and CLIP-Event). The loss function in Eq. (1) is extended by including similarity scores between the visual information and the negative prompts in the denominator.

### 3.4. Beyond VidSitu: Synthetic SRLs from Kinetics

We advocate the use of dense text prompts for efficient post-pretraining of CLIP for videos. Thus, while VidSitu is a natural fit, our work is *not* specific to VidSitu. While sparse labels generally perform worse (e.g. ViFi-CLIP), we present an approach to leverage action labels in Kinetics-700 and create dense SRL prompts that are effective for adaptation. Given video frames, we prompt LLaVA-1.6 (llava-hf/llava-v1.6-mistral-7b-hf) to generate role labels through questions: e.g. for the action label *driving*, we ask “who is driving?” or “what are they driving?”. Thus, we create  $K$ -SRL: 40k Kinetics videos with sparse action labels, augmented with dense and automatic (albeit noisy) SRL annotations. Unlike VidSitu, these are single event descriptions. Optionally, we can train SRL-CLIP with this data.

### 3.5. Implementation Details

We use the OpenAI CLIP implementation and its associated checkpoints [44], restricting experiments to ViT-B/32, ViT-B/16 and ViT-L/14 models. We add LoRA adapters to the CLIP image encoder and freeze both the CLIP image and text encoders. We find  $r=64$  rank to work well in our

Table 1. Zero-shot text-to-video retrieval on LSMDC and MSRVT. Metrics are recall at 5 (R5 $\uparrow$ ), at 10 (R10 $\uparrow$ ), and median rank (MdR $\downarrow$ ). Models in section B inherit CLIP, pretrained on 400M samples [44], and are further post-pretrained: CLIP4Clip on HowTo100M-380k [33, 35]; ViFi on Kinetics-400 [22]; and SRL-CLIP on VidSitu’s 23k videos [47]. See Sec. 4.1 for details.

Method	DSize	Params	LSMDC			MSRVT		
			R5	R10	MdR	R5	R10	MdR
<i>A. Non-CLIP based models</i>								
VideoCLIP [58]	1M	–	–	–	–	22.2	30.0	–
Frozen [3]	5M	232M	–	–	–	44.6	56.6	7
Clover [20]	5M	–	29.2	38.2	24	49.5	60.0	6
Singularity [24]	5M	209M	–	–	–	50.2	59.5	–
HiTeA [65]	5M	297M	31.1	39.8	–	54.2	62.9	–
ALPRO [25]	5.5M	231M	–	–	–	44.7	55.4	–
OmniVL [52]	18M	–	–	–	–	58.4	66.6	–
VideoCoCa [64]	100M	2.1B	–	–	–	57.8	67.0	–
VIOLET [11]	183M	198M	–	–	–	49.5	59.7	–
Florence [66]	900M	637M	–	–	–	63.8	72.6	–
<i>B. CLIP based models with simple post-pretraining</i>								
CLIP ViT-B/32	–	150M	28.9	35.7	31	53.2	63.0	4
CLIP ViT-B/16	–	150M	32.4	40.4	21	55.0	65.3	4
CLIP ViT-L/14	–	400M	36.8	43.7	19	59.5	69.6	<b>3</b>
CLIP4Clip [33]	0.4M	150M	28.5	36.4	28	57.0	66.9	4
ViFi [45] ViT-B/16	0.3M	150M	10.6	14.8	199	26.6	33.4	41
ViFi-IFT [45] ViT-B/16	0.3M	150M	32.6	39.9	25	57.6	67.1	<b>3</b>
SRL-CLIP ViT-B/32	<b>23k</b>	150M	31.1	39.2	24	58.2	69.4	<b>3</b>
SRL-CLIP ViT-B/16	<b>23k</b>	150M	35.7	43.8	17	59.7	71.8	<b>3</b>
SRL-CLIP ViT-L/14	<b>23k</b>	400M	<b>41.7</b>	<b>48.7</b>	<b>12</b>	64.9	73.6	<b>3</b>
<i>C. CLIP based models with sophisticated architecture/modality extensions</i>								
BT-Adapter [30]	2M	450M	35.9	45.0	–	64.7	73.5	–
ImageBind [14]	3M	1B	–	–	–	61.8	70.0	–
UMT-L [29]	5M	304M	37.2	43.7	–	58.1	66.7	–
TVTSv2 [67]	8.5M	1B	32.5	41.4	20	62.4	73.2	<b>3</b>
VAST [8]	154M	1.3B	–	–	–	<b>68.3</b>	<b>73.9</b>	–

experiments. Our VC module consists of 6 Tx encoder layers. We use  $\lambda=0.25$  for combining video- and event-level losses. The number of artificial hard negatives  $\mathcal{N}_{vr}=4$ . We use a learning rate of  $10^{-6}$  and the AdamW [32] optimizer. Each video in VidSitu has  $P=5$  events, and we sub-sample  $T=4$  frames from each event, for a total of  $L=20$  frames for a 10s video. We post-pretrain for 40 epochs on *one 12GB RTX2080 GPU* for ViT-B/32 & ViT-B/16 and *one 48GB RTX A6000 GPU* for ViT-L/14 with a batch size of  $B=20$  videos (100 event-text pairs).

## 4. Experiments

We evaluate SRL-CLIP on a variety of video understanding tasks that require different levels of perceptual granularity. Then, we present thorough ablations on the VidSitu dataset, providing insights about various design choices.

### 4.1. Zero-shot Text-to-Video (T2V) Retrieval

We present results on zero-shot T2V retrieval. To evaluate SRL-CLIP’s representations for coarse video understanding,

the VC is ignored here and retrieval scoring is performed by extracting frame-level features from SRL-CLIP’s backbone, followed by simple mean pooling, similar to [45].

We evaluate on two popular datasets: MSRVTT [59] and LSMDC [46]. Results obtained using standard retrieval metrics (recall and mean/median rank) are presented in Tab. 1. We also report post-pretraining efficiency by showing the dataset size and model size. We categorize and describe the competing methods in three sections:

**A. Non-CLIP based models** generally require large models and large datasets to perform on par with CLIP based models. Methods like VideoCoca [64] and Florence [66] are trained on 100M and 900M video samples and adopt ViT-H as a backbone, resulting in 2.1B and 637M parameters respectively. They require pretraining from scratch on multiple A100 GPUs for several days.

On the other hand, we build on top of CLIP (ViT-B/16 with 150M parameters and ViT-L/14 with 400M parameters), pretrained on 400M *images* (an order of magnitude smaller when compared to the number of *video frames* used above). Our post-pretraining (LoRA adaptation) on 23k video clips from VidSitu is highly efficient: it can be performed on a single 12GB RTX 2080 GPU within 5 hours for the base model, and on a 48GB A6000 GPU within 18 hours for the large model. On MSRVTT, our ViT-L/14 model outperforms VideoCoCa by 6.6% R@10 and surpasses Florence by 1.0% R@10 and outperforms all other models in this category (Tab. 1A).

**B. CLIP-based model with simple finetuning.** Our approach, SRL-CLIP, belongs to this category. CLIP4Clip uses a Transformer-based frame feature aggregator, and is post-pretrained on 0.4M videos while ViFi-IFT uses simple mean pooling, and is fine-tuned on 300k videos. Our approach, SRL-CLIP (ViT-L/14), fine-tuned on 23k videos with simple mean pooling, improves over CLIP4Clip and ViFi-IFT on all metrics. *E.g.* R@10 improves by 12.3% and 8.8% on LSMDC, and 6.7% and 6.5% on MSRVTT in comparison to CLIP4Clip and ViFi-IFT respectively. Fig. 3 (right) shows some qualitative results. We observe that SRL-CLIP is able to correctly understand fine-grained details such as red dress (example 1) or associate multi-person events across multiple shots (example 3) better than base CLIP.

**C. CLIP-based models with sophisticated extensions.** In this part, we compare against models with much larger sizes (*e.g.* BT-Adapter [30] at 450M, VAST [8] at 1.3B parameters) that are often trained with multiple modalities (*e.g.* ImageBind [14], VAST), with post-pretraining datasets 85–6700× larger than our 23k videos. SRL-CLIP (ViT-L/14) achieves strong results: (i) TVTSv2 [67] is worse on LSMDC (7.3% R@10) and is also slightly worse on MSRVTT (0.4% R@10). (ii) ImageBind [14], on MSRVTT, has worse R@10 by 3.6% and worse R@5

Table 2. ZS T2V retrieval performance on MSRVTT and LSMDC on the ViT-L/14 variants. VS = VidSitu, KS = Kinetics-SRL.

# Model	DSize	VS	KS	MSRVTT			LSMDC		
				R@5	R@10	MdR	R@5	R@10	MdR
1 CLIP	–	–	–	59.0	69.6	3	36.8	43.7	19.0
2 SRL-CLIP	23k	✓	–	64.9	73.6	3	<b>41.7</b>	48.7	11.5
3 SRL-CLIP	40k	–	✓	64.1	73.8	3	40.2	47.8	12.0
4 SRL-CLIP	63k	✓	✓	<b>65.0</b>	<b>74.0</b>	3	41.2	<b>50.8</b>	<b>10.0</b>

Table 3. Performance on VidSitu [47]. Action recognition measured as top-1 and top-5 verb accuracy (Vb@1, Vb@5). Captioning performance measured through CIDEr. We see a large performance improvement over base CLIP, while also achieving a new SoTA.

Method	Vb@1 ↑	Vb@5 ↑	CIDEr ↑
VidSitu [47]	46.79	75.90	46.01
Slow-D+TxE+TxD [57]	–	–	60.34
VideoWhisperer [23]	45.06	75.59	68.54
CLIP (ViT B/32)	44.68	79.79	55.14
SRL-CLIP (ViT B/32)	<b>45.88</b>	<b>80.66</b>	<b>71.50</b>
CLIP (ViT B/16)	45.83	80.12	54.25
SRL-CLIP (ViT B/16)	<b>48.78</b>	<b>81.95</b>	<b>72.11</b>
CLIP (ViT L/14)	46.57	78.73	60.79
SRL-CLIP (ViT L/14)	<b>52.36</b>	<b>83.91</b>	<b>76.24</b>

by 3.1%. (iii) UMT-L [29] is worse by 5.0% R@10 on LSMDC, and is worse by 6.9% R@10 on MSRVTT. Finally, SRL-CLIP outperforms BT-Adapter by 0.1% R@10 on MSRVTT. VAST performs marginally better than SRL-CLIP (0.3% R@10 on MSRVTT), however, both these models have higher model footprints (450M and 1.3B) and are post-pretrained on 2M and 154M videos (ours on 23k videos).

These results highlight the efficacy achievable through post-pretraining on dense and structured SRL, as opposed to narrations or captions.

**Kinetics-700.** Tab. 2 shows the results for adapting CLIP with K-SRL and VidSitu. Even though the K-SRL annotations are noisy, we improve R@5 by 5.1% for CLIP (row3 vs. row1) for MSRVTT while 3.4% for LSMDC. Training on K-SRL followed by VidSitu further improves over only training on VidSitu. Row4 vs. row2 shows a 0.8 reduction in MnR and a 0.4% increase in R@10 for MSRVTT. For LSMDC, there is a significant reduction in MnR (by 6.2) and a 2.1% increase in R@10.

## 4.2. Holistic Video Understanding

We evaluate SRL-CLIP’s holistic video understanding ability from two perspectives: (i) VidSitu [47] requires fine-grained perception to generate complex structured outputs; and (ii) Dense video captioning [6] requires localizing and describing actions in longer (few minute) videos.

**Video situation recognition** (VidSitu) [47] requires a model to predict the primary action verb in each event and generate

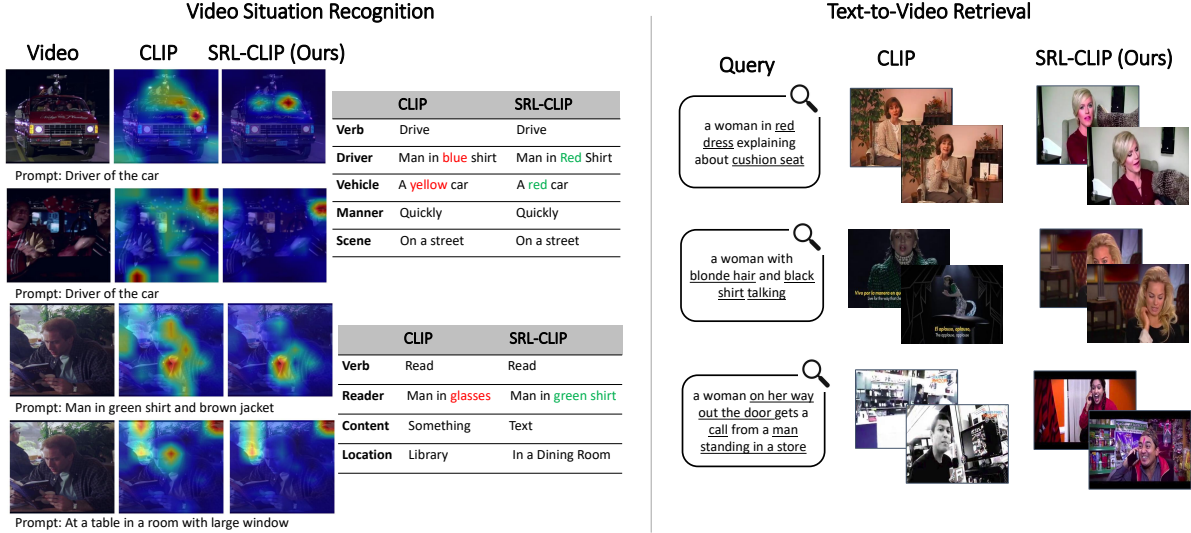


Figure 3. Qualitative results comparing CLIP and SRL-CLIP. **VidSitu** (left) shows improved attention maps resulting in better noun (SRL) captions; and **T2V Retrieval** (right) shows that SRL-CLIP has better awareness to details.

noun captions for the verb-appropriate set of roles. We adopt VideoWhisperer’s three-stage architecture [23] as it is the SoTA on VidSitu and remove the 165 object (Faster-RCNN) features and 5 event/action (SlowFast) features. Instead, as a simplification, we represent the  $P=5$  events as mean-pooled CLIP encoded frames.

In Table 3, we see that CLIP features perform on par with previous best methods on verbs (action understanding) achieving close to 45% Vb@1 accuracy. However, they perform much worse on CIDEr (55%) that evaluates quality of semantic role labels. This is expected as SRL requires fine-grained understanding that CLIP lacks [63].

Next, we evaluate SRL-CLIP by representing the  $P$  event features with event-level contextualized tokens (post VC). Verb accuracy improves by 5.8% for ViT L/14, while we see a large increase in SRL prediction performance with CIDEr improving by 15.5%. We acknowledge that a part of this gain may be due to the in-domain post-pretraining on VidSitu, especially training VC from scratch. Nevertheless, this demonstrates the effectiveness of our adaptation process and the video contextualizer that compresses information effectively. With this, we also set a new SoTA on VidSitu. Finally, in Fig. 3, we show that our adapted model is better suited for fine-grained reasoning with focused attention maps and accurate SRL prediction.

**VELOCITI** [48] is a recent benchmark used to test compositional reasoning abilities in video-language models. It has a series of 7 tests to evaluate different aspects of compositional reasoning for agents, actions, and event ordering. The results are shown in Tab. 4. SRL-CLIP outperforms Base CLIP in all tests except ActBind. Further, SRL-CLIP trained on VidSitu attains the highest accuracy on AgRand and ActMan, whereas when it is trained on Kinetics, it performs well on AgBind, ActAdv and AgCoref. Finally, SRL-CLIP trained

Table 4. Performance on the recent VELOCITI benchmark. We report the ClassicVLE accuracy across 7 different tests on the ViT-L/14 variant. VS = VidSitu, KS=Kinetics-SRL.

Method	Data	Ag	Ag	Act	Act	Act	Ag
		Rand	Bind	Adv	Man	Bind	Coref
CLIP	–	82.47	56.07	64.84	54.80	<b>58.48</b>	53.06
SRL-CLIP	VS	<b>87.97</b>	55.17	63.93	<b>60.26</b>	55.01	55.16
SRL-CLIP	KS	85.80	<b>56.27</b>	<b>65.75</b>	56.77	57.52	<b>57.82</b>
SRL-CLIP	VS+KS	87.17	52.98	64.38	58.95	55.09	54.28

on both VidSitu and Kinetics obtains slight improvements on the Chrono test.

### 4.3. Ablations

We evaluate the impact of various design choices, with results on VidSitu or MSRVT. All ablations are performed on the CLIP ViT L/14 model.

**How to adapt CLIP?** We start with identifying how and which layers of the backbone should be adapted. Tab. 5 shows various options from full (F) or partial fine-tuning (P), and using LoRA modules (L) on the image encoder (IE) and text encoder (TE). Column 2 vs. 3 and 4 vs. 5 show that freezing TE improves performance. We suspect this is because it encourages IE to align with fine-grained descriptions. Column 5 shows good SRL CIDEr performance with LoRA for IE and frozen TE. Comparing columns 1 and 5 we see that full fine-tuning (F) is not only costly, but also performs worse than LoRA (1.1% Vb@1 and 1% CIDEr), likely due to concept forgetting. We adopt LoRA fine-tuning for the IE with a frozen TE as our default configuration.

**Impact of loss functions** is presented in Tab. 6. Directly adapting the CLIP backbone with the CLIP-Event and CLIP-Video losses (R1-2) results in poor SRL performance on VidSitu. The benefits of a video contextualizer (VC), both

Table 5. Ablation on CLIP adaptation strategy. F: Full fine-tuning, P: Partial fine-tuning, L: LoRA, IE/TE: Image/Text encoder.

Backbone	IE TE	F -	P P	P -	L L	L -
VidSitu	Vb@1 CIDEr	43.53 70.27	42.83 67.83	42.87 71.00	<b>45.53</b> 70.14	44.65 <b>71.27</b>

Table 6. Ablation on loss functions. Results on VidSitu. CE: CLIP-Event, CV: CLIP-Video, VCE: VC-Event, VCV: VC-Video.

	Losses				ViT-L/14		ViT-B/32	
	CE	CV	VCE	VCV	Vb@1	CIDEr	Vb@1	CIDEr
1	✓	-	-	-	52.78	59.08	44.77	58.37
2	✓	✓	-	-	51.93	58.27	<b>45.38</b>	57.34
3	-	-	✓	-	<b>52.72</b>	75.21	43.75	71.19
4	-	-	✓	✓	52.08	75.03	45.12	70.02
5	✓	-	✓	✓	52.66	73.66	44.46	<b>72.10</b>
6	✓	✓	✓	✓	52.36	<b>76.24</b>	44.65	71.27

Table 7. Ablation study on when should VC be used. PPT: Post-pretraining, DT: downstream-task.

(a) VidSitu (Video Situation Recognition)						
	VC		ViT-B/32		ViT-L/14	
	PPT	DT	Vb@1	CIDEr	Vb@1	CIDEr
1		CLIP	44.7	55.1	46.6	60.8
2	-	-	45.4	57.3	51.9	58.3
3	✓	-	<b>46.7</b>	59.8	<b>53.1</b>	63.4
4	✓	✓	45.9	<b>71.5</b>	52.4	<b>76.2</b>

(b) MSRVT (zero-shot T2V retrieval)										
	VC		ViT-B/32				ViT-L/14			
	PPT	DT	R@5	R@10	Mn.R	Md.R	R@5	R@10	Mn.R	Md.R
1	CLIP		53.2	63.0	41.2	4	59.0	69.6	33.6	3
2	-	-	57.5	68.1	29.5	4	64.3	<b>75.7</b>	<b>21.7</b>	<b>2</b>
3	✓	-	<b>58.2</b>	<b>69.4</b>	<b>27.8</b>	<b>3</b>	<b>64.7</b>	73.8	22.5	3
4	✓	✓	15.9	24.7	120.2	46	32.8	43.5	71.4	15

during training and downstream evaluation, are seen in R4 that uses VC-Event and VC-Video losses, achieving good results on VidSitu SRL (R4: 75.0% vs. R2: 58.3% CIDEr for ViT-L/14). R5-6 combine both CLIP and VC losses. We choose R6 as the default model as it achieves the highest geometric mean (good trade-off) between Vb@1 and CIDEr scores, yielding 52.36% Vb@1 and 76.24 CIDEr.

**When should VC be used?** We investigate the need for VC and its influence on downstream results in Tab. 7. We evaluate on two tasks, in-domain VidSitu and zero-shot T2V retrieval on MSRVT.

First, row 1 (R1) shows results for the original CLIP model without adaptation, copied here for ease of comparison. In R2, we directly use CLIP-Video and CLIP-Event losses; while R3 uses VC only for post-pretraining (PPT). In all three cases, downstream tasks (DT) are performed by extracting features directly from the CLIP backbones.

For ViT-B/32, we observe improvements across all met-

rics as we step from R1 to R2 to R3. For zero-shot T2V on MSRVT, we see that R2 (direct CLIP PPT) brings a large improvement over R1 (5.1% R@10). This highlights the benefit of using structured SRL for efficient CLIP video adaptation. Nevertheless, including VC results in a further 1.3% improvement on R@10 (R3). On VidSitu, we see steady improvements on CIDEr: 2.2% from R1 to R2, and 2.5% from R2 to R3.

Next, we analyze what happens when using VC in both the post-pretraining and the downstream task (R4), again for ViT-B/32. Specifically, we use the  $P$  contextualized event representations  $\hat{e}_{ik}$  for VidSitu and the single contextualized video representation  $\hat{v}_i$  for MSRVT. The VC helps compress and contextualize multiple events in a video, leading to strong performance improvements on VidSitu (SRL CIDEr improves by 11.7% over R3). As VC is trained only on VidSitu, it learns to accommodate the complexity of VidSitu. For example, the overall video representation is trained to match against a dense description of multiple events. However, when VC is used directly on MSRVT, the performance collapses as the captions in MSRVT are not as descriptive.

We observe similar overarching trends for the larger ViT-L/14 model, where VC significantly boosts fine-grained VidSitu performance (a 12.8% CIDEr improvement in R4 over R3), though for coarse retrieval on MSRVT, direct post-pretraining without VC (R2) is sufficient to achieve peak performance.

## 5. Conclusion

We proposed a recipe for efficient and effective post-pretraining of CLIP to adapt it to videos, using a densely annotated video-caption dataset. Specifically, we curated template-based prompts from a video SRL dataset [47], that captures visual concepts in a video holistically using action verbs, person-object nouns, attributes, locations, *etc.* We hypothesized that post-pretraining on dense prompts provides a strong learning signal to the visual encoder, which not only allows for highly efficient but also effective adaptation to videos. Our adapted model SRL-CLIP, exhibited superior or comparable performance on zero-shot text-to-video retrieval against SOTA models having 4–8× more parameters and post-pretrained on up to 6000× more data. Additionally SRL-CLIP shows good downstream generalization, consistently outperforming CLIP on multiple video benchmarks requiring different levels of perceptual granularity. Specifically, we achieve SoTA performance on a fine-grained and holistic video understanding benchmark VidSitu [47].

**Acknowledgments.** The project was supported by funding from SERB SRG/2023/002544. MT also thanks support from the Google India Faculty Award.

**Limitations.** SRL-CLIP has some limitations to address in

future work:

(1) As seen in Tab. 7, the VC may work well only on tasks similar to the PPT dataset. However, the VC alone cannot solve the PPT task and the CLIP backbone is indeed improved, evident through our downstream experiments.

(2) The VidSitu dataset comprises movie clips, that are person-centric. Thus, we expect SRL-CLIP to work well broadly on videos with people. Please note that VidSitu is used as an example in our work, and our take-away message is that *dense text prompts can efficiently improve representations*.

(3) Our paper proposes that PPT on a small, densely annotated dataset is effective. However, we acknowledge that curating such a dataset is challenging and costly. While recent VLM advances (e.g. [56]) may offset some of these annotations, the upfront annotation costs offset massive computation costs later.

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 23716–23736, 2022. 1
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4
- [3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *International Conference on Computer Vision (ICCV)*, pages 1728–1738, 2021. 1, 2, 3, 5
- [4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A CLIP-Hitchhiker’s Guide to Long Video Retrieval. *arXiv preprint arXiv:2205.08508*, 2022. 3
- [5] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the “Video” in Video-Language Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2917–2927, 2022. 2
- [6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 6
- [7] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3
- [8] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 5, 6, 2
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 4
- [10] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. CLIP2Video: Mastering Video-Text Retrieval via Image CLIP. *arXiv preprint arXiv:2106.11097*, 2021. 2, 3
- [11] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. VIOLET: End-to-End Video-Language Transformers with Masked Visual-token Modeling. *arXiv preprint arXiv:2111.12681*, 2021. 5
- [12] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal Transformer for Video Retrieval. In *European Conference on Computer Vision (ECCV)*, pages 214–229, 2020. 2
- [13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-

- adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 3
- [14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 5, 6
- [15] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5006–5015, 2022. 3
- [16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The” something something” video database for learning and evaluating visual common sense. In *International Conference on Computer Vision (ICCV)*, pages 5842–5850, 2017. 3
- [17] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient Low Rank Adaptation of Large Models. *arXiv preprint arXiv:2402.12354*, 2024. 3
- [18] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. In *International Conference on Machine Learning (ICML)*, pages 2790–2799, 2019. 3
- [19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3
- [20] Jingjia Huang, Yinan Li, Jiashi Feng, Xinglong Wu, Xiaoshuai Sun, and Rongrong Ji. Clover: Towards a Unified Video-Language Alignment and Fusion Model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14856–14866, 2023. 5
- [21] Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*, 2022. 3
- [22] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natssev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5
- [23] Zeeshan Khan, CV Jawahar, and Makarand Tapaswi. Grounded Video Situation Recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:8199–8210, 2022. 6, 7
- [24] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing Single Frame Bias for Video-and-Language Learning. *arXiv preprint arXiv:2206.03428*, 2022. 5
- [25] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and Prompt: Video-and-Language Pre-Training With Entity Prompts. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4953–4963, 2022. 5
- [26] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1
- [27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022.
- [28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*, 2023. 1
- [29] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*, 2023. 5, 6
- [30] Ruyang Liu, Chen Li, Yixiao Ge, Ying Shan, Thomas H Li, and Ge Li. One for all: Video conversation is feasible without video instruction tuning. *arXiv preprint arXiv:2309.15785*, 2023. 1, 5, 6, 2
- [31] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-Decomposed Low-Rank Adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 3
- [32] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 5
- [33] 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. *arXiv preprint arXiv:2104.08860*, 2021. 1, 2, 3, 5
- [34] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM Multimedia (MM)*, pages 638–647, 2022. 2, 3
- [35] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2630–2640, 2019. 1, 2, 3, 5
- [36] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *International Conference on Computer Vision (ICCV)*, pages 15579–15591, 2023. 1
- [37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*, 2018. 3
- [38] Junting Pan, Ziyi Lin, Xiatao Zhu, Jing Shao, and Hongsheng Li. ST-Adapter: Parameter-Efficient Image-to-Video Transfer Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 26462–26477, 2022. 3

- [39] Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisín Mac Aodha. SVL-Adapter: Self-Supervised Adapter for Vision-Language Pretrained Models. In *British Machine Vision Conference (BMVC)*, 2022. 3
- [40] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations (ICLR)*, 2021. 2
- [41] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Disentangling Spatial and Temporal Learning for Efficient Image-to-Video Transfer Learning. In *International Conference on Computer Vision (ICCV)*, pages 13934–13944, 2023. 3
- [42] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Disentangling Spatial and Temporal Learning for Efficient Image-to-Video Transfer Learning. In *International Conference on Computer Vision (ICCV)*, pages 13934–13944, 2023. 2
- [43] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019. 2
- [44] 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 (ICML)*, pages 8748–8763. PMLR, 2021. 1, 3, 5
- [45] Hanoona Rasheed, Muhammad Uzair khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned CLIP models are efficient video learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 5, 6
- [46] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision (IJCV)*, 123:94–120, 2017. 6
- [47] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual Semantic Role Labeling for Video Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5589–5600, 2021. 1, 2, 3, 5, 6, 8
- [48] Darshana Saravanan, Varun Gupta, Darshan Singh, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. VELOCITI: Benchmarking Video-Language Compositional Reasoning with Strict Entailment. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 7
- [49] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5227–5237, 2022. 3
- [50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 4
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3
- [52] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. OmniVL: One Foundation Model for Image-Language and Video-Language Tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:5696–5710, 2022. 5
- [53] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Action-CLIP: A New Paradigm for Video Action Recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2, 3
- [54] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual Expert for Pretrained Language Models, 2023. 1
- [55] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified Coarse-to-Fine Alignment for Video-Text Retrieval. *arXiv preprint arXiv:2309.10091*, 2023. 2
- [56] Wenliang Dai and Junnan Li and Dongxu Li and Anthony Meng Huat Tiong and Junqi Zhao and Weisheng Wang and Boyang Li and Pascale Fung and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 9
- [57] Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. Hierarchical self-supervised representation learning for movie understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9727–9736, 2022. 6
- [58] 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. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6787–6800, 2021. 2, 5
- [59] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 6
- [60] Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. Pixel Aligned Language Models. *arXiv preprint arXiv: 2312.09237*, 2023. 1
- [61] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5036–5045, 2022. 2, 3
- [62] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Alignment. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3
- [63] Yutaro Yamada, Yingtian Tang, and Ilker Yildirim. When are Lemons Purple? The Concept Association Bias of CLIP. *arXiv preprint arXiv:2212.12043*, 2022. 7

- [64] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. VideoCoCa: Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners. *arXiv preprint arXiv:2212.04979*, 2022. [5](#), [6](#)
- [65] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. HiTeA: Hierarchical Temporal-Aware Video-Language Pre-training. In *International Conference on Computer Vision (ICCV)*, pages 15405–15416, 2023. [5](#)
- [66] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. [5](#), [6](#)
- [67] Ziyun Zeng, Yixiao Ge, Zhan Tong, Xihui Liu, Shu-Tao Xia, and Ying Shan. TVTsv2: Learning Out-of-the-box Spatiotemporal Visual Representations at Scale. *arXiv preprint arXiv:2305.14173*, 2023. [5](#), [6](#), [2](#)

# SRL-CLIP: Efficient CLIP Video Adaptation via Structured Semantic Role Labels

## Supplementary Material

Table 8. Impact of varying the no. of verb-role hard negatives,  $\mathcal{N}_{vr}$ .

$\mathcal{N}_{vr}$	VidSitu	
	Vb@1	CIDEr
0	52.38	75.63
1	51.72	74.78
2	52.54	74.02
3	53.26	74.88
4	<b>52.36</b>	<b>76.24</b>

Table 9. Impact of adapting different weights using LoRA. q, k, v, and o are the query, key, value, and output projection matrices in the self-attention block. fc and proj are the two MLPs after the self-attention module.

Weight type	VidSitu	
	Vb@1	CIDEr
q k v	<b>52.36</b>	<b>76.24</b>
q k v o	<b>52.62</b>	73.13
q k v o fc	51.22	73.61
q k v o proj	51.83	71.86

## Appendix

We present additional results and discussions in the supplementary material. Sec. 6 starts with additional ablation experiments, focusing on hard negatives (HNs) and the low-rank adaptation (LoRA) module. In Sec. 7 we present qualitative results on the various tasks shown in the main paper. This provides an opportunity to better understand the working of SRL-CLIP. All ablation studies are conducted using the ViT-L/14 variant.

## 6. Additional Ablations

**How many Hard Negatives to use?** Tab. 8 shows the impact of varying the number of hard negatives (HNs). Here, we see that as  $\mathcal{N}_{vr}$  increases, the verb prediction accuracy increases while CIDEr drops by a small amount. This is expected as verb-role HNs tend to improve verb prediction performance. The best performance is considered at the geometric mean between Vb@1 and CIDEr on the VidSitu task.

**What is the best LoRA configuration?** In Tab. 9, we study the impact of adapting different weights of the SRL-CLIP image encoder with LoRA. We get the best performance when we include LoRA modules only for the attention weights

Table 10. ZS T2V on MSRVT with CLIP ViT B/32. We increase the amount of VidSitu training data, resulting in a consistent improvement in performance for R@10 and MnR.

Data	VidSitu	
	Vb@1	CIDEr
0%	46.57	60.79
50%	50.82	70.72
70%	50.62	73.31
90%	51.36	73.55
100%	<b>52.36</b>	<b>76.24</b>

in the Transformer (“q k v” –  $W_q, W_k, W_v$ ) while keeping everything else frozen. Hence, in our default model, we only adapt the self-attention matrices (parameters) with LoRA.

**Training on subsets of VidSitu.** The results of increasing the post-pretraining dataset from 10% to 100% are presented in Tab. 10. The table shows a consistent decline in mean rank and an improvement in R@10 with additional data. Additionally, with just 10% of the data, a mere 2,300 videos, we see a 3.9% boost in R@5 and a 2.6% in R@10; indicating the effectiveness of dense SRL prompts.

**Detailed descriptions incur labeling costs, but save on compute costs.** The use of dense prompts results in efficient training that completes in 5 hours on a *single* RTX2080 GPU (12 GB). In contrast, large models adapted on large (noisy) datasets incur significantly higher compute costs requiring many and larger GPUs. *E.g.* BT-Adapter [30] uses 8 V100 GPUs, TVTSv2 [67] uses 80 V100 GPUs, and VAST [8] uses 64 V100 GPUs. Considering the typical number of hyperparameter evaluations and experiments required, improving data quality is not only cheaper but also more viable in the long-term.

**More examples of natural and artificial hard negatives** are shown in Tab. 11. Notice how the naturally occurring hard negatives are good enough to learn strong video representations even without the need for artificial hard negatives. Different from most works that only use text-based negatives, VidSitu also facilitates visual natural negatives for the same text. Note how the hard negative captions are very plausible; in example 1 the action *look* instead of *speak*.

## 7. Qualitative Results

We now present qualitative results on 6 datasets. When not mentioned otherwise, we use the default variant of SRL-CLIP.

Positive Prompt	Natural Hard Negatives	Verb-role Hard Negatives
In this photo, the <i>action</i> is <b>speaking</b> where, the <i>talker</i> is <b>man standing in yellow sweatshirt</b> , the <i>hearer</i> is <b>woman with scarf</b> , the <i>manner</i> is <b>standing in the middle of a full airplane</b> , the <i>scene</i> of the event is <b>an airplane</b> .	In this photo, the <i>action</i> is <b>turn</b> where, the <i>the turner</i> is <b>man standing in yellow sweatshirt</b> , the <i>the thing turning</i> is <b>his body</b> , the <i>direction</i> is <b>towards woman with scarf</b> , the <i>scene</i> of the event is <b>an airplane</b> .	In this photo, the <i>action</i> is <b>look</b> where, the <i>looker</i> is <b>man standing in yellow sweatshirt</b> , the <i>thing looked at</i> is <b>woman with scarf</b> , the <i>direction</i> is <b>is to his back</b> , <i>manner</i> is <b>standing in the middle of a full airplane</b> , the <i>scene</i> of the event is <b>an airplane</b> .
In this photo, the <i>action</i> is <b>open</b> where, the <i>opener</i> is <b>man in brown jacket and man in gray suit</b> , the <i>the thing opening</i> is <b>trunk of taxi</b> , the <i>manner</i> is <b>annoyed</b> , the <i>scene</i> of the event is <b>near a taxi</b> .	In this photo, the <i>action</i> is <b>hoist</b> where, the <i>lifter</i> is <b>man in brown jacket and man in gray suit</b> , the <i>thing going up</i> is <b>dead body</b> , the <i>direction</i> is <b>up into trunk of taxi</b> , the <i>scene</i> of the event is <b>near a taxi</b> .	In this photo, the <i>action</i> is <b>respond</b> where, the <i>replier</i> is <b>man in brown jacket and man in gray suit</b> , the <i>scene</i> of the event is <b>near a taxi</b> .
In this photo, the <i>action</i> is <b>bow</b> where, the <i>bower</i> is the <b>woman in glasses</b> , the <i>bowed to</i> is <b>man wearing black</b> , the <i>manner</i> is <b>on her knees</b> , the <i>scene</i> of the event is <b>in a well lit room</b> .	In this photo, the <i>action</i> is <b>photograph, take a picture</b> where, the <i>photographer</i> is the <b>man in black suit</b> , the <i>subject</i> is <b>woman in glasses</b> , the <i>scene</i> of the event is <b>in a well lit room</b> .	In this photo, the <i>action</i> is <b>smash</b> where, the <i>smasher</i> is the <b>woman in glasses</b> , the <i>smashed</i> is <b>man wearing black</b> , the <i>direction</i> is <b>on patients face</b> , the <i>scene</i> of the event is <b>in a well lit room</b> .

Table 11. We show the naturally occurring hard negatives in a batch as well as the process of converting a standard positive prompt into hard negatives by swapping verb-role information. The template is shown in gray, e.g. In this photo,. The action and roles are shown in italics, e.g. *action, talker; hearer*. The correct prompt values (verbs or nouns) are in cobalt blue, e.g. *speaking, man standing in yellow sweatshirt*; and the replaced verbs, roles, or nouns are in deep red. We swap the verb and roles in verb-role hard negatives while keeping the same nouns and performing some mapping between previous and new roles.

**MSRVTT.** We show zero-shot text-to-video retrieval on the MSRVTT dataset in Fig. 4. We can see that SRL-CLIP performs much better than CLIP when the queries have a compositional nature. The last row shows a failure case. Although SRL-CLIP retrieves a video in which a man is talking, he is not talking about hiking. It is hard to pick the right video just by using the visual modality, as hiking is not very clear by just watching the video. In fact, SRL-CLIP retrieves a video shot outdoors, which may have been associated with hiking rather than the indoor video.

**LSMDC.** We show zero-shot text-to-video retrieval on the LSMDC dataset in Fig. 5. LSMDC is a much harder dataset compared to MSRVTT, as it is based on movies that contain more dynamic shot changes. Also, the agent/patient of an action is annotated as “SOMEONE”, unlike VidSitu, where they are described according to their characteristics, making it even more challenging. We can see that SRL-CLIP outperforms CLIP here as well.

**VidSitu.** Fig. 6 shows the qualitative results on video situa-

tion recognition for 5 videos. SRL-CLIP outperforms CLIP, especially when picking attributes like color. SRL-CLIP is also better at predicting the role *manner* (which captures the expression/emotion of the person), which CLIP struggles with. However, both SRL-CLIP and CLIP show similar (good) performance when predicting the *scene*. The last row shows a failure case (note that CLIP also fails to give good noun captions in this case). It is interesting to see that SRL-CLIP correctly identifies the *reacher* as a boy but assigns the wrong attribute to it.

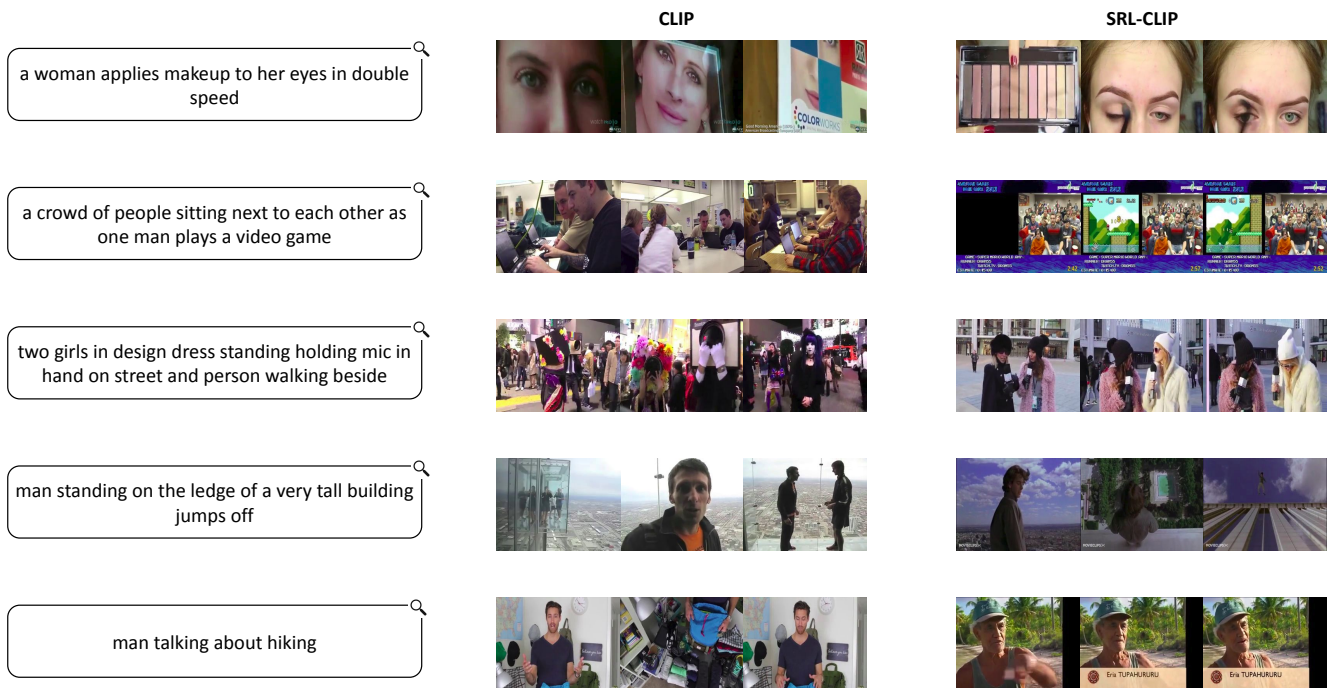


Figure 4. Zero-shot text-to-video retrieval on the MSRVTT dataset. We show three frames of the top-1 retrieved video for each query. We can see that SRL-CLIP outperforms CLIP, specially when compositional reasoning is required. The last row shows a failure case. Although SRL-CLIP retrieves a video in which a man is talking, and potentially with more appropriate background, he is not talking about hiking.



Figure 5. Zero-shot text-to-video retrieval on the LSMDC dataset. We show three frames of the top-1 retrieved video for each query. We can again notice that SRL-CLIP performs better than CLIP when compositional reasoning is needed. The last row shows a failure case.



	verb	driver	vehicle	direction	manner	scene
CLIP	drive	man in hat	car	down the street	slowly	in a car
SRL-CLIP	drive	man in black jacket	car	down the road	with a serious look on his face	in a car
GT	drive	man in a black police uniform	car	forward	intently	car

	verb	entity entering	thing entered	manner	scene
CLIP	walk	man in blue shirt	the house	with his right hand	outside a house
SRL-CLIP	enter	boy in red shirt	the door	slowly	outside a building
GT	enter	kid in red	door	casually	in doorway

	verb	looker	looked at	direction	manner	scene
CLIP	stare	woman with dark hair	man in green shirt	down	with a slight smile	in a room
SRL-CLIP	stare	woman with blonde hair	man in brown jacket	down	with a sad expression	in a room
GT	look	girl with blonde hair	a man in front of her	forward	sadly	in a room

	verb	talker	hearer	manner	scene
CLIP	look	girl with dark hair	man in blue shirt	while seated next to each other	in a room
SRL-CLIP	speak	woman in a green shirt	man in brown shirt	while face to face	in a kitchen
GT	speak	blonde girl	old man	while sitting down	cabin

	verb	reacher	body part	goal	direction	purpose	scene
CLIP	kneel	woman in blue coat	hand	to grab something	down	to get something	in a room
SRL-CLIP	kneel	boy in orange shirt	his body	to grab something	down	to pick up a plate	in a room
GT	grab	boy	his hand	booklet	towards the man	to take booklet from man	lab

Figure 6. Video Situation Recognition on 5 videos. SRL-CLIP performs much better than CLIP in picking the right attribute of an entity. The last row shows a failure case where the semantic role labels predicted by SRL-CLIP deviates from the ground-truth (GT).