CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Alignment

Hongwei Xue\textsuperscript{1}; Yuchong Sun\textsuperscript{2}; Bei Liu\textsuperscript{2}; Jianlong Fu\textsuperscript{3};
Ruihua Song\textsuperscript{2}; Houqiang Li\textsuperscript{1}; Jiebo Luo\textsuperscript{4}
\textsuperscript{1}University of Science and Technology of China, Hefei, China,\n\textsuperscript{2}Renmin University of China, Beijing, China,\n\textsuperscript{3}Microsoft Research, Beijing, China,\n\textsuperscript{4}University of Rochester, Rochester, NY

Abstract
Pre-trained image-text models, like CLIP, have demonstrated the strong power of vision-language representation learned from a large scale of web-collected image-text data. In light of the well-learned visual features, there are works that transfer image representation to the video domain and achieve good results. However, adapting image-text pre-trained models to video-text pre-training (i.e., post-pretraining) has not demonstrated a significant advantage yet. In this paper, we tackle this challenge by raising and addressing two questions: 1) what are the factors hindering post-pretraining CLIP from improving performance on video-text tasks, and 2) how to mitigate the impact of these factors. Through a series of comparative experiments and analyses, we find that the data scale and domain gap between language sources have large impacts. By these observations, we propose an Omnisource Cross-modal Learning method equipped with a Video Proxy mechanism on the basis of CLIP, namely CLIP-ViP. Extensive results show that our approach improves the performance of CLIP on video-text retrieval by a large margin. Our model achieves state-of-the-art results on a variety of datasets, including MSR-VTT, DiDeMo, LSMDC, and ActivityNet. We release our code and pre-trained CLIP-ViP models at https://github.com/microsoft/XPretrain/tree/main/CLIP-ViP.

1 Introduction

In the past few years, vision-language pre-training has achieved great success on cross-modal representation learning from a large scale of web-crawled data (Radford et al., 2021; Jia et al., 2021; Li et al., 2021; Wang et al., 2021b; Zellers et al., 2021; 2022; Bain et al., 2021). Among them, image-text pre-trained models (Radford et al., 2021; Jia et al., 2021) have shown powerful capability for various downstream tasks, including visual understanding (Gu et al., 2021; Wang et al., 2021a; Rao et al., 2022), image-text generation (Patashnik et al., 2021; Mokady et al., 2021) and so on (Guzhov et al., 2022; Zhang et al., 2022). In light of the well-learned and enriched visual representation, some works directly adapt image-text pre-trained models to video-text downstream tasks without further pre-training on video data (Luo et al., 2021; Fang et al., 2021; Gorti et al., 2022; Zhao et al., 2022), while still outperforming models pre-trained on video data (Xu et al., 2021b; Bain et al., 2021).

Utilizing an existing powerful image-text pre-trained model for further video-text pre-training (i.e., post-pretraining) is able to reduce the required training cost by making good use of the knowledge learned from images. However, adapting image-text pre-trained models to video-text data for post-pretraining has not demonstrated a significant advantage yet, and thus is still under-explored. A preliminary study is conducted by CLIP4Clip (Luo et al., 2021) which adopts MeanPooling by averaging multiple frame features based on the CLIP model on a subset of Howto100M (Miech et al., 2019). While the improvement over directly using the image-text pre-trained model is marginal for

\textsuperscript{*}Equal contribution. This work was performed when Hongwei Xue and Yuchong Sun were visiting Microsoft Research as research interns.

\textsuperscript{†}Corresponding authors.
either zero-shot or fine-tuning settings. In this paper, we aim to explore how to effectively adapt the image-text pre-trained model (e.g., CLIP) to video-language representation learning for video-text tasks (e.g., text-to-video retrieval).

To unleash the power of video data to adapt image-text pre-trained models for post-pretraining, we conduct several preliminary experiments to figure out the challenges that hinder post-pretraining. First, we explore post-pretraining an image-text pre-trained model (i.e., CLIP) with MeanPooling on video-text datasets with different scales, including WebVid-2.5M (Bain et al., 2021) and HD-VILA-100M (Xue et al., 2022). The result shows that the scale of data is critical for video-text post-pretraining. Data on a small scale makes the model easy to over-fit the new data while the knowledge learned from image-text is suppressed and the performance is reduced. Second, we investigate the language domain gap between pre-training data and downstream data. By calculating the Normalized Mutual Information (NMI) on clusters of text features, we find that there is a large domain gap between subtitles that are used in large-scale video-text pre-training data and descriptive texts in downstream tasks.

To mitigate the impact of the above factors, we propose CLIP-ViP to adapt the pre-trained image-text model CLIP for video-text pre-training. First, we introduce auxiliary captions that have a smaller language domain gap with downstream data into existing large-scale video-text data. Instead of using a video captioning model which may cause data leakage by training on the same dataset with video-text downstream tasks, and considering a better visual captioning capability, we adopt an image captioning model to generate an auxiliary caption of middle frame in each video. In order to adapt a Transformer-based vision encoder to process both images and videos with minimal modification, we then propose video proxy tokens and design a proxy-guided video attention mechanism for the Vision Transformer (ViT). Specifically, during attention computation in each block, video proxy tokens can interact with all tokens, while patch tokens only interact with video proxy tokens and patch tokens within the same frame. Our vision encoder only increases negligible parameters and calculations compared to the vanilla Vision Transformer while increasing the generality and extendability. To facilitate cross-modal representation learning from both caption-frame and video-subtitle data types at the same time, we propose an Omnisource Cross-modal Learning (OCL) method for pre-training and study a series of variants to find the best fusion strategy.

Our experimental results show that our approach improves the performance of CLIP on text-to-video retrieval tasks by a large margin. We also conduct ablation studies to verify the effectiveness of each part in our approach. Our contributions are summarized as follows: (1) We are one of the first to explore factors that hinder video post-pretraining on pre-trained image-text models; (2) We propose CLIP-ViP that can effectively leverage image-text pre-trained model for post-pretraining; (3) We conduct extensive experiments to verify the effectiveness of our method. Our model outperforms the state-of-the-art results by a large margin on four widely-used benchmarks.

2 RELATED WORK

Vision-Language Pre-Training End-to-end models (Lei et al., 2021; Xue et al., 2022; Zellers et al., 2021; Fu et al., 2021; Huang et al., 2020; 2021b; Xue et al., 2021; Li et al., 2021; Kim et al., 2021; Huang et al., 2021a; Sun et al., 2022) for vision-language pre-training are replacing the traditional approach using pre-extracted visual features by off-the-shelf models (Sun et al., 2019; Xu et al., 2021b; Zhu & Yang, 2020; Li et al., 2020b; Chen et al., 2020). Training end-to-end models on large-scale web-collected data also gradually demonstrates the big advantages (Radford et al., 2021; Jia et al., 2021; Xue et al., 2022; Zellers et al., 2021; 2022). Unlike images that have alt-texts, large-scale video datasets suitable for pre-training usually use subtitles as text sources (Miech et al., 2019; Xue et al., 2022). Subtitles are much noisier than alt-texts, according to (Miech et al., 2019), typical examples of incoherence include the content producer asking viewers to subscribe to their channel, talking about something unrelated to the video, or describing something before or after it happens. Bain et al. collect a video dataset WebVid (Bain et al., 2021) with textual description annotations. Their texts are well aligned with the video and avoid suffering from ASR errors. However, the vast majority of WebVid videos are sourced from a stock footage website, so scaling up is under limitation. The video-subtitle data is more easily accessible on the web and thus suitable for scaling up. In this paper, we investigate the unfavorable factors of video-subtitle data and explore how to mitigate the impact of these factors.
Pre-trained Models for Video-Text Retrieval  

The great success of the CLIP has demonstrated its unprecedented power on varies downstream tasks, including vision understanding (Gu et al., 2021; Wang et al., 2021a; Rao et al., 2022), image-text generation (Patashnik et al., 2021; Mokady et al., 2021) and so on (Guzhov et al., 2022; Zhang et al., 2022). By contrastive learning on large-scale image-text pairs, CLIP learns enriched visual concepts for images. Recently, some works directly transfer CLIP to video-text retrieval without further pretraining on video data (post-pretraining) (Luo et al., 2021; Fang et al., 2021; Gorti et al., 2022; Zhao et al., 2022; Wang et al., 2022c). Their work takes the performance of video-text retrieval to a new level, outperforming existing models pre-trained on video data (Xu et al., 2021b; Bain et al., 2021; Xue et al., 2022; Ge et al., 2022; Wang et al., 2022c). They transfer CLIP from views of feature aggregation (Luo et al., 2021; Zhao et al., 2022; Fang et al., 2021; Gorti et al., 2022) or representation alignment (Fang et al., 2021; Gorti et al., 2022; Wang et al., 2022c). In parallel with these works, we study post-pretraining with video data on top of CLIP in an effective way and our model can be combined with other approaches effectively.

3  Preliminary Analysis

In this section, we first study the impact of the data scale for adapting image-text pre-training to video-text post-pretraining, and then investigate how the language domain gap affects the adaption.

3.1  Post-pretraining with Different Data Scales

To study the effectiveness of different data scales, we use the CLIP-ViT-B/32 model (Radford et al., 2021) as the base image-text pre-trained model and adopt MeanPooling for video adaption like CLIP4Clip (Luo et al., 2021) by averaging multiple frame features as video feature. Two open-domain video-text datasets are used: WebVid-2.5M (Bain et al., 2021) with 2.5 million pairs and HD-VILA-100M (Xue et al., 2022) with 100M pairs. We also adopt a subset of HD-VILA-100M containing random 10% data (namely HD-VILA-10M) as a middle setting. We run the same number of steps on all settings, equivalent to one epoch on HD-VILA-100M. We uniformly sample 12 frames from each video and apply the same hyper-parameters as described in Section 5 for all settings.

During post-pretraining, we evaluate the pre-trained models by fine-tuning on MSR-VTT text-to-video retrieval task. Figure 1 shows the performance trend. We observe an overfitting phenomenon that continuous post-pretraining leads to a performance drop. And the drop is more significant for smaller data (e.g., WebVid-2.5M and HD-VILA-10M). As CLIP is pre-trained on 400 million image-text pairs, further training on small data makes the model tend to overfit the new data while the implicit knowledge learned from the image-text pairs is degrading. As a consequence, the performance will drop, even worse than using CLIP directly. Thus we adopt HD-VILA-100M due to its large scale and diverse category.

3.2  Language Domain Gap with Downstream Data

It is intuitive that pre-training on data with the same domain as downstream data can benefit downstream tasks. For most video-text tasks like video-text retrieval, texts are descriptive sentences of videos (i.e., captions). While for HD-VILA-100M, which we will use for pre-training, the texts are auto-transcribed subtitles and they indicate very different relevance to visual information compared to descriptive texts. Meanwhile, auto-transcribed subtitles suffer from irrelevance, misalignment, and ASR errors (Miech et al., 2019). To better explore the language domain gap between pre-
4 APPROACH

In this section, we will introduce the proposed CLIP-ViP video pre-training framework. To bridge language domain gaps between image and video datasets, we first introduce an in-domain auxiliary data generation method. Then, we propose a novel Video Proxy mechanism to enable the Vision Transformer (ViT) model for both image and video encoding. We further present an Omnisource Cross-modal Learning (OCL) method which can jointly learn cross-modal representation from video-text and image-text pairs.

4.1 IN-DOMAIN AUXILIARY DATA GENERATION

Motivated by the analysis in Section 3, we introduce auxiliary captions into large-scale video-subtitle data to reduce the language domain gap between pre-training and downstream data. We adopt an
image captioning model for two reasons. 1) Most SOTA video captioning models are trained with video-text datasets (e.g., MSR-VTT, ActivityNet) which are also used for downstream tasks. We avoid data leakage to perform pre-training agnostic to downstream data. 2) The performance of existing video captioning models lags far behind that of images. Thus, we choose a powerful image captioning model OFA-Caption (Wang et al., 2022b) to generate one caption for the middle frame of each video in HD-VILA-100M. We use the default setting of the OFA-Caption model. As a result, we generate 100M sentences with a max length of 16 words. This method can be applied to any video data and we will release the generated captions to facilitate future research.

4.2 Video Proxy Mechanism

Since video is an ordered sequence of frames, it is critical to learn the frame aggregation and temporality when transferring to the video domain. Meanwhile, to keep the high generality and extendability of the Vision Transformer (ViT) backbone, we aim to find a simple but effective way to transfer ViT to enable both image and video encoding with minimal modifications. Given a video containing \( T \) frames: \( \{f_1, f_2, ..., f_T\} \), we follow CLIP to divide each frame into \( N \) patches: \( \{f^1_t, f^2_t, ..., f^N_t | t \in [1, T]\} \). Then we add spatio-temporal positional embedding to each flattened 2D patches:

\[
g(f^n_t) = \text{Linear}(f^n_t) + \text{Pos}_s(n) + \text{Pos}_t(t),
\]

where \( \text{Linear}(*) \) is a linear layer, \( \text{Pos}_s(n) \) and \( \text{Pos}_t(t) \) is the learnable spatial and temporal positional embedding, respectively. The whole video can be divided into \( T \times N \) patch tokens.

To model spatial information from multi-frames, one simple way is directly feeding all tokens into CLIP’s vision encoder and conducting attention across all tokens. However, this method introduces significant conflicts with CLIP. As CLIP is pre-trained on image and text pairs, it has difficulty handling interactions of tokens between images/frames during training. We also verify it by experiments as Full Attention setting in Table 2. Instead, we introduce a Video Proxy token to act as a proxy that will be regarded as the video’s representation. We also design a proxy-guided attention mechanism for the vanilla ViT. In the attention score calculation of each block, video proxy tokens attend to all other frame tokens, while patch tokens only attend to tokens in the same frame plus video proxy tokens. This mechanism can be formulated as attention mask \( \mathcal{M}_{\text{ViP}} \):

\[
\mathcal{M}_{\text{ViP}}(q, k) = \begin{cases} 
1 & \text{if } q \in \mathcal{P} \text{ or } k \in \mathcal{P} \text{ or } (q, k) \text{ in the same frame,} \\
0 & \text{otherwise},
\end{cases}
\]

where \( q \) and \( k \) is the key and query in attention calculation. Patch tokens can obtain global information from video proxy tokens while reducing inconsistencies with the original CLIP’s calculation. Our experiment in Section 5 demonstrates the superiority of this mechanism.

For the input type of the image/frame, we use linear interpolation to get a middle temporal positional embedding, then treat the image/frame as a special single-frame video. This method enables joint training on both videos and images in the same batch, as our proxy-guided attention mechanism reduces the difference in calculations between video and image.

4.3 Omnisode Cross-Modal Learning

To learn rich video-language alignment from video-subtitle pairs and reduce the language domain gap with downstream data by corresponding auxiliary frame-caption pairs, we study joint Cross-Modal Learning on the omnisode input. Following most works of learning multimodal alignment on dual encoders (Radford et al., 2021; Xue et al., 2022; Li et al., 2021; Luo et al., 2020; Xu et al., 2021b; Luo et al., 2021), we use info-NCE loss to perform contrastive learning. There are two formats of visual source: video sequences and single frames, and two types of text source: subtitles and captions in our work. We denote them by \( V, F, S, \) and \( C \) respectively. We define a source-wise info-NCE loss by:

\[
\begin{align*}
\mathcal{L}_{v2t} &= -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{v_i^T t_i / \tau}}{\sum_{j=1}^{B} e^{v_j^T t_i / \tau}}, \\
\mathcal{L}_{t2v} &= -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{t_i^T v_i / \tau}}{\sum_{j=1}^{B} e^{t_j^T v_i / \tau}}
\end{align*}
\]
We study the reasonable variants of OCL: (a) $\mathcal{L}_{V \leftrightarrow S} + \mathcal{L}_{F \leftrightarrow C}$: Simple combination of two source-wise losses on video-subtitle and frame-caption pairs; (b) $\mathcal{L}_{V \leftrightarrow S} + \mathcal{L}_{V \leftrightarrow C}$: As there is also content correlation between videos and its middle-frame captions, we explore to add a loss on video-caption pairs to baseline loss $\mathcal{L}_{V \leftrightarrow S}$; (c) $\mathcal{L}_{V \leftrightarrow S} + \mathcal{L}_{V \leftrightarrow C} + \mathcal{L}_{F \leftrightarrow C}$: Combination of (a) and (c); (d) $\mathcal{L}_{V \leftrightarrow S, C} + \mathcal{L}_{F \leftrightarrow C}$: A video corresponds to both a subtitle and auxiliary caption. Compare to (c), the numbers of negative pairs in $\mathcal{L}_{V \leftrightarrow C}$ can be expanded. The $\mathcal{L}_{V \leftrightarrow C}$ in $\mathcal{L}_{V \leftrightarrow S, C}$ is rewritten as:

$$
\mathcal{L}_{V \leftrightarrow C} = \frac{1}{2B} \sum_{i=1}^{B} \left( \log \frac{e^{v_i^T s_i / \tau}}{\sum_{j=1}^{B} e^{v_i^T s_j / \tau} + e^{v_i^T c_j / \tau}} + \log \frac{e^{v_i^T c_i / \tau}}{\sum_{j=1}^{B} e^{v_i^T c_j / \tau} + e^{v_i^T s_j / \tau}} \right),
$$

where $s_i \in S$ and $c_i \in C$. The $\mathcal{L}_{V \leftrightarrow S, C}$ in $\mathcal{L}_{V \leftrightarrow S}$ is equal to (e). We compare all variants with the baseline $\mathcal{L}_{V \leftrightarrow S}$ and report results in Section 5.

## 5 Experiment

### 5.1 Experimental Details

**Video-Text Post-Pretraining.** To pre-train the proposed CLIP-ViP model, we uniformly sample 12 frames and resize all frames to $224 \times 224$ from video clips with an average length of 13.4 seconds. The sampled frames can well cover the semantics conveyed from a video clip. For text, we adopt the CLIP’s tokenizer to split a sentence into word tokens with a max length of 70. We use AdamW optimizer (Loshchilov & Hutter, 2019), and empirically set an initial learning rate as $5 \times 10^{-6}$ and a fixed weight decay as $5 \times 10^{-2}$. For the learning rate schedule, we adopt a cosine decay with a warm-up strategy. We train our model with 32 NVIDIA Tesla V100 GPUs in a batch size of 1024. The contrastive similarity is calculated on gathered features from all GPUs. We set training steps to one epoch on HD-VILA-100M for all ablation studies and three epochs for the full setting.

**Fine-tuning Training.** To better adapt CLIP-ViP to downstream tasks, we reuse most hyper-parameters of post-pretraining in fine-tuning with some exceptions: 1) Batch size: we fine-tune on CLIP-ViT-B/32. All post-pretraining steps are equivalent to one epoch on HD-VILA-100M.
Table 4: Ablation study of post-pretrain data. We report text-to-video results of models finetuned on MSR-VTT and DiDeMo. Mean ↑ indicates an average of Recall@1, 5 and 10. For all results, the model is designed with 4 video proxy tokens and pre-trained on CLIP-ViT-B/32. All post-pretraining steps are equivalent to one epoch on HD-VILA-100M.

<table>
<thead>
<tr>
<th>Post-pretrain Data</th>
<th>MSR-VTT Retrieval</th>
<th>DiDeMo Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1 ↑</td>
<td>R@5 ↑</td>
</tr>
<tr>
<td>w/o Post-pretrain</td>
<td>46.5</td>
<td>72.1</td>
</tr>
<tr>
<td>HD-VILA_{sub}</td>
<td>47.7</td>
<td>72.1</td>
</tr>
<tr>
<td>HD-VILA_{cap}</td>
<td>45.9</td>
<td>73.0</td>
</tr>
<tr>
<td>HD-VILA_{sub+cap}</td>
<td>49.6</td>
<td>74.5</td>
</tr>
<tr>
<td>ImageCaption</td>
<td>45.6</td>
<td>70.7</td>
</tr>
<tr>
<td>HD-VILA_{sub} + ImageCaption</td>
<td>49.1</td>
<td>73.1</td>
</tr>
</tbody>
</table>

Omnisource Cross-modal Learning. To verify the effectiveness of the proposed Omnisource Cross-modal Learning (OCL) and compare its variants, we set a post-pretraining and fine-tuning pipeline and adopt the same hyper-parameters for all experiments. $\mathcal{L}_{Y+\delta}$ is the baseline contrastive...
loss on video-subtitle pairs. After introducing auxiliary captions, we study four variants of OCL Loss: (a) $L_{V\leftrightarrow S} + L_{F\leftrightarrow C}$; (b) $L_{V\leftrightarrow S} + L_{V\leftrightarrow C}$; (c) $L_{V\leftrightarrow S} + L_{V\leftrightarrow C} + L_{F\leftrightarrow C}$; (d) $L_{V\leftrightarrow S, C} + L_{F\leftrightarrow C}$ as explained in Section 4.3. We pre-train models with each loss function for only one epoch due to the costly training, then finetune on two video-text retrieval datasets: MSR-VTT and DiDeMo. We compare the results with CLIP-MeanPool and CLIP using the proposed Video Proxy mechanism without post-pretraining (i.e., CLIP-Vip). The results are listed in Table 3. On MSR-VTT dataset, we find that $L_{V\leftrightarrow S}$ brings very little improvement: 0.4% on average of Recall@1, 5, 10. This is due to the large domain gap between MSR-VTT and post-pretraining data. Combined with auxiliary captions, four variants of OCL loss all bring significant improvements: over 3% on Recall@1 and over 2.3% on average of Recall@1, 5, 10. On DiDeMo dataset, based on the improvement brought by $L_{V\leftrightarrow S}$, OCL further improve the results by a large margin: 8% on average of Recall@1, 5, 10. Finally, $L_{V\leftrightarrow S, C} + L_{F\leftrightarrow C}$ performs best which is applied as our final setting.

**Auxiliary Data.** In this part, we ablate the contribution of large-scale noisy data and auxiliary data. For uni-source, we use video-subtitle pairs and video-caption data for post-pretraining by vanilla contrastive loss. For data combination, we apply OCL under $L_{V\leftrightarrow S, C} + L_{F\leftrightarrow C}$ setting to post-pretrain on the combined data. From Table 4, Omnisource post-pretraining results are much better than two uni-source results. On MSR-VTT, both uni-source post-pretraining show limited improvement: 67.4% and 66.9% compared with 67.0%. While the Omnisource post-pretraining brings a significant improvement of 2.6%. On DiDeMo, the benefit of data combination is also obvious, with nearly double the improvements brought by uni-source. These results show that the auxiliary data together with our designed joint learning method can effectively adapt the image-text pre-trained model to video-text post-pretraining. As the generation of auxiliary captions is based on OFA-Caption (Wang et al., 2022b), a powerful image-text pre-trained model, we also explore only including existing data in post-pretraining. We choose image-text pairs of several widely-adopted datasets: MS-COCO, Visual Genome (VG) (Krishna et al., 2017b), Flickr-30K (Young et al., 2014), SBU (Ordonez et al., 2011), CC3M (Sharma et al., 2018) and CC12M as our auxiliary data (namely ImageCaption). To ablate the contribution of these data, we add experiments of post-pretraining on ImageCaption alone and HD-VILA-100M.
5.3 COMPARISON TO STATE-OF-THE-ART MODELS

We compare our model under full setting (in three epochs) with state-of-the-art works on the text-to-video retrieval task. The results of fine-tuning on four datasets (i.e., MSR-VTT, DiDeMo, ActivityNet Captions, LSMDC) are shown in Table 5 and 6, respectively. We clarify the backbone for CLIP-based works. We only add results with DSL (Cheng et al., 2021) to make fair comparison with some methods using post-processing operations e.g., DSL (Cheng et al., 2021) or QB-Norm (Bogolin et al., 2022). Our model achieves the best results on all datasets in both CLIP-ViT-B/32 and CLIP-ViT-B/16. Note that some existing methods are also applicable on top of our models as our modification to the CLIP model is minimal. Note that even without post-processing (e.g., DSL), our results still surpass methods using post-processing operations on most datasets. Besides, adding DSL will greatly improve the performance of our model since our model has good bidirectional vision-language correspondence. The good results on the ActivityNet Captions dataset also indicate that our model can generalize well to long videos. Overall, the improvements on different datasets demonstrate the superiority of the video-language representation learned by our CLIP-ViP model.

6 CONCLUSION

In this paper, we study further pre-training (post-pretraining) image-text models like CLIP on large-scale video data. We first conduct a preliminary analysis to reveal the factors hindering video post-pretraining. Motivated by findings, we propose CLIP-ViP which includes an Omnisource Cross-modal Learning method and a Video Proxy mechanism. The Video Proxy mechanism can better model videos containing temporal information while reducing conflicts with the pre-trained CLIP model. The Omnisource Cross-modal Learning alleviates the problem caused by the domain gap between video-subtitle and downstream data. Extensive results show that our approach improves the performance of CLIP on video-text retrieval by a large margin and also achieves new state-of-the-art results on four widely-used video-language benchmarks.
REFERENCES


Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In ICCV, 2021. 1, 2, 3, 8, 9


Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In CVPR, pp. 3558–3568, 2021. 4


Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In CVPR, pp. 16167–16176, 2022. 3, 8, 9


Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In ACM MM, pp. 1138–1147, 2021a. 2


Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In CVPR, pp. 12976–12985, 2021b. 2
Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pp. 4904–4916. PMLR, 2021. 1, 2


Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pp. 706–715, 2017a. 7, 8


Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pp. 11336–11344, 2020a. 2


Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 7


Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pp. 2630–2640, 2019. 1, 2, 3


Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 24, 2011. 8


A MORE DOWNSTREAM TASKS

We add various suitable downstream tasks to verify that our model can be generalizable to other tasks. Specifically, we add four zero-shot evaluations on video action recognition and video-language multi-choice tasks:

**HMDB51 Action Recognition**  HMDB51 (Kuehne et al., 2011) is a video dataset containing realistic videos from various sources, including movies and web videos. There are 51 action categories (such as “jump”, “kiss” and “laugh”). We evaluate our model by zero-shot classification on HMDB51. To adapt to our model, we formulate video classification as video-text alignment, using text prompts “a person [label]”. The prediction is made on top-1 similarity between query video and prompted texts. We directly test our model without finetuning to report zero-shot accuracy results.

**SomethingSomethingV2 (SSv2) Action Recognition.** SomethingSomethingV2 (Goyal et al., 2017) dataset is a large collection of labeled video clips that show humans performing pre-defined basic actions with everyday objects. It requires models to develop fine-grained understanding of basic actions that occur in the physical world. To adapt to our model, we design a multi-choice task on SSv2. The model has to choose the ground truth annotation among all categories. We directly test our model without finetuning to report zero-shot accuracy results.

**MSR-VTT Multi-Choice (MC).** MSR-VTT MC is a benchmark build on MSR-VTT video dataset. Given a video query and five descriptive sentences, MSR-VTT MC requires the model to choose a single best answer in the candidates. We directly test our model without finetuning to report zero-shot accuracy results.

**LSMDC Multi-Choice (MC).** LSMDC MC is a benchmark build on movie data. Similar with MSR-VTT MC, it requires the model to choose a single best answer from 5 candidates. We report zero-shot accuracy on 10000 test videos.

From the results in Table 7, we can find that post pretraining CLIP on subtitle alone will lead to a performance drop on most tasks: HMDB51 SSv2 and MSR-VTT MC. This phenomenon is in consistency with our assumption as subtitles usually have unique form. By our final setting, CLIP-ViP
achieves better results than CLIP on all evaluations, especially on SSv2 which requires more understanding on temporality (Wang et al., 2022d). These results demonstrate that CLIP-ViP outperforms CLIP on various video benchmarks and further show the generalization of our model CLIP-ViP.

### B Efficiency of Post-pretraining

In light of the good image representation of CLIP, our post-pretraining significantly reduces the training cost. Compared with HD-VILA (Xue et al., 2022) which trains video-language representation from scratch, CLIP-ViP outperforms HD-VILA by a large margin with about 1/30 training time. To make more fair comparison with training from scratch. We conduct an experiment of training from scratch with the same setting and training time (one epoch). Then we finetune on MSR-VTT Retrieval to evaluate the pre-trained model.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSR-VTT R@1</th>
<th>MSR-VTT R@5</th>
<th>MSR-VTT R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP-ViP (from scratch)</td>
<td>28.5</td>
<td>55.2</td>
<td>68.6</td>
</tr>
<tr>
<td>CLIP-ViP</td>
<td>49.6</td>
<td>74.5</td>
<td>84.8</td>
</tr>
</tbody>
</table>

Table 8: Comparison with training CLIP-ViP from scratch. We report finetuning results on MSR-VTT Retrieval. All models are base size with patch size of 32. Details of post-pretraining are the same as in Section 5.2.

From the results in Table 8, we can see that there is a huge performance gap. Without basing on a CLIP, CLIP-ViP can not leverage the rich knowledge entailed in CLIP, thus leading to a much lower training efficiency.