Unifying Latent and Lexicon Representations for Effective Video-Text Retrieval

Haowei Liu[1,2*], **Yaya Shi**[3*], **Haiyang Xu**[4†], **Chunfeng Yuan**[1†], **Qinghao Ye**[4]
Chenliang Li[4], **Ming Yan**[4], **Ji Zhang**[4], **Fei Huang**[4], **Bing Li**[1], **Weiming Hu**[1,2,5]
[1]MAIS, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]University of Science and Technology of China [4]Alibaba Group
[5]School of Information Science and Technology, ShanghaiTech University, China
liuhaowei2019@ia.ac.cn, shiyaya@mail.ustc.edu.cn
{shuofeng.xhy, ym119608}@alibaba-inc.com, {cfyuan, bli, wmhu}@nlpr.ia.ac.cn

Abstract

In video-text retrieval, most existing methods adopt the dual-encoder architecture for fast retrieval, which employs two individual encoders to extract global latent representations for videos and texts. However, they face challenges in capturing fine-grained semantic concepts. In this work, we propose the UNIFY framework, which learns lexicon representations to capture fine-grained semantics and combines the strengths of latent and lexicon representations for video-text retrieval. Specifically, we map videos and texts into a pre-defined lexicon space, where each dimension corresponds to a semantic concept. A two-stage semantics grounding approach is proposed to activate semantically relevant dimensions and suppress irrelevant dimensions. The learned lexicon representations can thus reflect fine-grained semantics of videos and texts. Furthermore, to leverage the complementarity between latent and lexicon representations, we propose a unified learning scheme to facilitate mutual learning via structure sharing and self-distillation. Experimental results show our UNIFY framework largely outperforms previous video-text retrieval methods, with 4.8% and 8.2% Recall@1 improvement on MSR-VTT and DiDeMo respectively. Code and pre-trained models will be publicly available at https://github.com/auhowielau/UNIFY.

Keywords: video-text retrieval, lexicon representation, unified learning

1. Introduction

Video-text retrieval is a crucial task with wide practical applications. Recently, pre-training to learn transferable cross-modal representations has gradually become the paradigm of this field (Bain et al., 2021; Li et al., 2022b; Bai et al., 2022; Wang et al., 2022; Ge et al., 2022a,b). To achieve fast retrieval, most methods adopt the dual-encoder architecture. It employs two individual encoders for video and text feature extraction respectively, and uses contrastive learning for cross-modal alignment.

As dual-encoder models compress a video (or text) into a latent vector, cross-modal interaction and alignment are solely based on such coarse-grained global representations. Therefore, it's challenging for them to capture fine-grained semantic concepts such as objects and actions. To tackle this, some methods (*e.g.* Li et al., 2022b) employ extra interaction modules to enhance global latent representations. However, it deprives the model's ability of fast retrieval. BridgeFormer (Ge et al., 2022a) discards the extra module after training, and thus the fine-grained interaction ability cannot be well transferred to the model when inference.

In this work, we present a novel **UNIFY** framework for unified video-text retrieval. It learns lexicon representations of videos and texts to capture fine-grained semantics, and combines the strengths of latent and lexicon representations for effective cross-modal retrieval. Firstly, we define a lexicon space where each dimension corresponds to a semantic concept represented by a word. As shown in Figure 1, videos and texts are mapped into this space to obtain lexicon representations. To

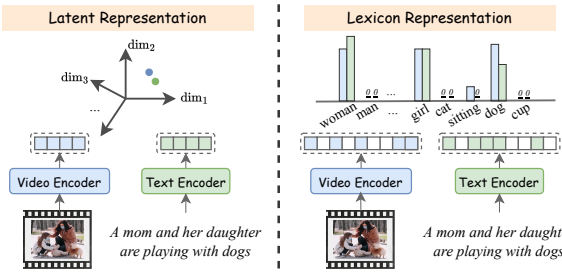

Figure 1: Comparison of latent and lexicon representations. The dimensions of latent representations have no explicit meanings. In contrast, each dimension of lexicon representations corresponds to a semantic concept, where semantically relevant dimensions are activated (*e.g.* woman and dog) while semantically irrelevant dimensions are suppressed (*e.g.* cat and cup).

*Equal contribution.
†Corresponding authors.

LREC-COLING 2024, pages 17031–17041

capture fine-grained semantic information, we propose a two-stage semantics grounding approach to activate semantically relevant dimensions and suppress semantically irrelevant dimensions. Secondly, as latent representations summarize videos and texts from a global perspective, and lexicon representations excel at capturing fine-grained semantics, combining them can further improve the model's performance. Thus, to better leverage their complementarity, we propose a unified learning scheme which facilitates mutual learning between them via structure sharing and self-distillation.

Specifically, inspired by SPLADE (Formal et al., 2021b), we can ground texts to semantically relevant dimensions by resorting to a pre-trained BERT (Devlin et al., 2018) model and its masked language modeling (MLM) head. However, it's much more intractable for videos to achieve this due to the giant gap between raw pixels and the lexicon space. To address this, we propose a two-stage semantics grounding approach. As initially videos have random distributions in the lexicon space, in stage one, we freeze the text encoder to avoid textual lexicon representations being corrupted in cross-modal alignment. We map local video and text features into the lexicon space using the MLM head, and aggregate them to obtain video-level and text-level lexicon representations. Contrastive learning is then applied to pull paired samples closer and push unpaired ones away. In stage two, we jointly train both the video and text encoders for further cross-modal alignment. Apart from video-text contrastive learning, we employ the MLM task in this stage to preserve textual semantics.

To leverage the complementarity between latent and lexicon representations, we propose a unified learning scheme. Firstly, from a structure sharing perspective, the two types of representations share a stem video (or text) encoder in shallow layers to promote knowledge sharing and transfer. Meanwhile, representation-specific encoders are adopted in deep layers to focus on global and fine-grained semantic information respectively. Secondly, as learning lexicon representations is relatively more challenging, we utilize latent representations to provide additional supervision information from a different perspective via self-distillation. Specifically, we employ the similarity scores computed from latent representations as soft labels for the contrastive learning of lexicon representations. Through the proposed unified learning scheme, latent and lexicon representations can benefit from each other, and are unified to form an effective video-text retriever.

Experimental results demonstrate the proposed lexicon representations can capture fine-grained semantics effectively. Moreover, our UNIFY framework combines the strengths of latent and lexicon representations, and largely outperforms previous state-of-the-art methods in video-text retrieval.

Our contributions can be summarized as follows:

- We present a novel UNIFY framework which unifies global latent representations and fine-grained lexicon representations for effective video-text retrieval.
- We propose a two-stage semantics grounding approach to ground videos and texts into semantically relevant dimensions, and a unified learning scheme to leverage the complementarity of latent and lexicon representations.
- Experimental results show our model well captures fine-grained semantics and largely surpasses previous video-text retrieval methods.

2. Related Work

2.1. Video-Text Retrieval

Recently, pre-training to learn transferable cross-modal representations has been popular in both image-text retrieval (Xu et al., 2021a; Li et al., 2022a; Xu et al., 2023) and video-text retrieval (Miech et al., 2020; Bain et al., 2021; Xu et al., 2021b; Ge et al., 2022a,b). However, dual encoder methods have shortcomings in understanding the fine-grained alignment between video and text, which is crucial for accurate video-text retrieval. Currently, there are two ways to solve this problem. The first (Li et al., 2022b; Ge et al., 2022a) involves using an additional fusion encoder to model fine-grained cross-modal interactions. However, the features cannot be pre-cached in this kind of structure and limit the model's fast retrieval ability. The second one takes some learning strategies based on the dual encoder, such as MILES (Ge et al., 2022b) introduces masked image modeling task to inject the fine-grained semantics into global representation. Nevertheless, the fine-grained semantics are learned in an implicit way which may not be the optimal approach. We believe that an explicit fine-grained representation will facilitate better retrieval performance. Therefore, in this paper, we introduce an efficient and effective method UNIFY by introducing a specific representation branch - lexicon one to capture fine-grained semantics, and we use a unifying scheme to promote collaboration between the fine-grained lexicon and original global latent representation.

2.2. Lexicon Representation

The concept of lexicon representation was initially introduced by Vector Space Model (Salton et al., 1975), which represents a document as a vector in a vector space. Each dimension corresponds

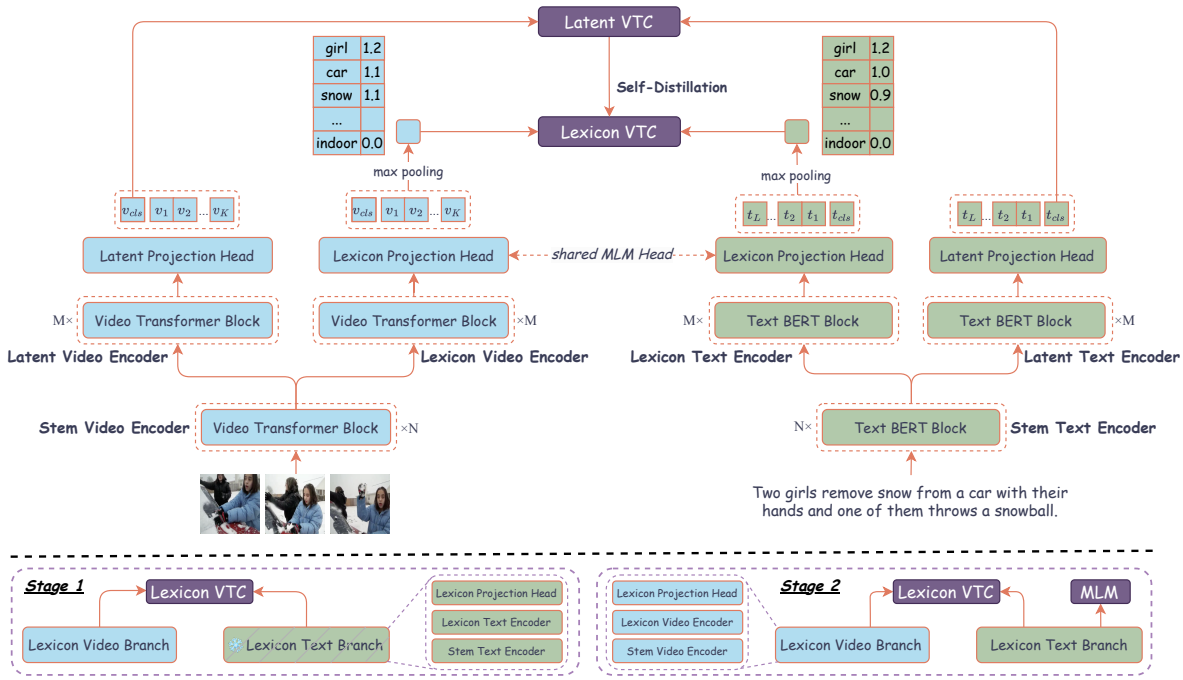

Figure 2: Overview of our proposed UNIFY framework. The whole model consists of two streams for video and text respectively, each including a stem encoder, two representation-specific encoders and two projection heads. For lexicon representation learning, we propose a two-stage semantics grounding approach (Section 3.2). Furthermore, we unify the latent and lexicon representations via structure sharing and self-distillation (Section 3.3). VTC stands for video-text contrastive learning.

to a term in the vocabulary, whose value indicates the importance of the term in the document. As pre-trained language models have gained popularity, neural network-based lexicon methods (Bai et al., 2020; Formal et al., 2021b,a; Lassance and Clinchant, 2022; Shen et al., 2022) have achieved progress. Our method is inspired by the neural-based lexicon representation methods. However, our task involves both video and text modalities. Though text data can be naturally projected into the lexicon space by directly taking the pre-trained language model as the text encoder, it's much more intractable for videos to achieve this due to the giant modality gap between video and text. To tackle this, in this paper, we propose a two-stage semantics grounding approach to learn lexicon representations of videos.

3. Method

3.1. Overview

As Figure 2 shows, we propose the UNIFY framework to unify latent and lexicon representations for effective video-text retrieval. The architecture of UNIFY can be roughly divided into two parts, *i.e.* the video stream and the text stream, to extract video and text representations respectively. Both parts consist of five elements, *i.e.*, a stem encoder $E_{stem}(\cdot)$ shared by the latent and lexicon representations, two representation-specific encoders $E_{lat}(\cdot)$, $E_{lex}(\cdot)$ and two corresponding projection heads $H_{lat}(\cdot)$, $H_{lex}(\cdot)$. Note that videos and texts share the same lexicon projection head to map both modalities into the same lexicon space.

It takes two steps to extract latent and lexicon representations from videos and texts. Taking an input video V for example, as the left part of Figure 2 shows, **1)** we first use the stem video encoder to extract video features, which are then fed into the latent video encoder and lexicon video encoder to obtain the corresponding raw features:

$$r^v_{lat} = E^v_{lat}(E^v_{stem}(V)) \in \mathbb{R}^d, \quad (1)$$

$$r^v_{lex} = E^v_{lex}(E^v_{stem}(V)) \in \mathbb{R}^{K\times d}, \quad (2)$$

where K is the number of local video features, and d is the dimension of raw features. Note that the latent encoder only outputs the global [CLS] feature, while the lexicon encoder outputs all local raw features. **2)** We employ the latent and lexicon projection heads to map raw features into the corresponding representation spaces. For latent representations, this step can be formulated as:

$$f^v_{lat} = H^v_{lat}(r^v_{lat}) \in \mathbb{R}^{\hat{d}}, \quad (3)$$

where $\hat{d}$ is the dimension of the latent space. As for the lexicon space, each dimension of it corresponds to a semantic concept represented by a

word. Denoting the lexicon size as $|\mathbb{W}|$, the lexicon projection can be formulated as:

$$p_{lex}^{v} = H_{lex}^{v}(r_{lex}^{v}) \in \mathbb{R}^{K\times|\mathbb{W}|}. \tag{4}$$

After getting the lexicon representations of local patches, we perform the $\mathrm{ReLU}(\cdot)$ activation function to suppress negative values to zero, and aggregate them to get the video-level lexicon representation by pooling operation:

$$f_{lex}^{v} = \mathrm{Pool}(\mathrm{ReLU}(p_{lex}^{v})) \in \mathbb{R}^{|\mathbb{W}|}. \tag{5}$$

As the right part of Figure 2 shows, following a similar process, we obtain the latent and lexicon representation of an input text T:

$$f_{lat}^{t} = H_{lat}^{t}(E_{lat}^{t}(E_{stem}^{t}(T))) \in \mathbb{R}^{\hat{d}}, \tag{6}$$

$$p_{lex}^{t} = H_{lex}^{t}(E_{lex}^{t}(E_{stem}^{t}(T))) \in \mathbb{R}^{L\times|\mathbb{W}|}, \tag{7}$$

$$f_{lex}^{t} = \mathrm{Pool}(\mathrm{ReLU}(p_{lex}^{t})) \in \mathbb{R}^{|\mathbb{W}|}, \tag{8}$$

where L is the number of tokens of T.

After obtaining the latent and lexicon representations of videos and text, when inference, we utilize both types of representations to calculate the dot product similarity scores (S_{lat} and S_{lex}) between video-text pairs. Finally, we combine the two scores at a 1:1 ratio to serve as the final similarity score for ranking:

$$S = S_{lat} + S_{lex}. \tag{9}$$

3.2. Two-stage Semantics Grounding

In order to capture fine-grained semantics, lexicon representations are required to satisfy two semantic constraints: **1)** dimensions semantically relevant to the video (or text) are activated with high values, and **2)** semantically irrelevant dimensions are suppressed to zero. This makes the learning of lexicon representations quite challenging. Luckily, for texts, we can resort to pre-trained language models (PLM) to achieve this goal. PLMs such as BERT (Devlin et al., 2018) are trained with the masked language modeling (MLM) task, which can project the masked tokens to semantically relevant words. Therefore, by reusing PLM and its MLM head, we can transform texts into lexicon representations that satisfies the semantic constraints. However, it's much more intractable for videos to achieve the same goal, as video's raw pixels are significantly different from discrete words in both modality and semantics. To tackle this, we propose a two-stage semantics grounding approach.

Stage 1. As initially videos are randomly distributed in the lexicon space, updating video and text encoders simultaneously may damage the lexicon distributions of texts. Therefore, as the lower left part of Figure 2 shows, we freeze the text encoders and ground videos into the lexicon space in stage 1. We resort to the paired texts to acquire the information which dimensions of the lexicon space are relevant to the videos. Specifically, we learn cross-modal semantic alignment in the lexicon space by optimizing a video-text contrastive (VTC) learning objective with Noise-Contrastive Estimation (NCE):

$$\begin{aligned}\mathcal{L}_{\mathrm{VTC}}^{lex} &= \sum_{i=1}^{B}\mathrm{NCE}\left(v_i, t_i\right) + \sum_{i=1}^{B}\mathrm{NCE}\left(t_i, v_i\right), \\ \mathrm{NCE}\left(x_i, y_i\right) &= -\log\frac{\exp\left(x_i^T y_i/\tau\right)}{\sum_{j=1}^{B}\exp\left(x_i^T y_j/\tau\right)},\end{aligned} \tag{10}$$

where v_i and t_i are normalized lexicon representations of the i-th video and text in a batch. B is batch size and τ is a temperature hyper-parameter.

Stage 2. As the lower right part of Figure 2 shows, in stage 2, we jointly train the video and text encoders for bidirectional cross-modal alignment. However, simply unfreezing the text encoders will deprive the semantic constraints and cause texts to drift in the lexicon space. Therefore, apart from video-text contrastive learning, we adopt the masked language modeling (MLM) task in stage 2 for preserving textual semantics. MLM recovers the masked tokens by reasoning contextual text, and thus encourages the model to project text tokens to lexicon dimensions corresponding to their semantics. Denoting the text with tokens masked as $\hat{T}$ and the prediction probability of the masked tokens as $\boldsymbol{p}^{\mathrm{mask}}(\hat{T})$, MLM loss can be formulated as:

$$\mathcal{L}_{\mathrm{MLM}} = \mathbb{E}_{\hat{T}\sim D}\left[\mathrm{CE}\left(\boldsymbol{y}^{\mathrm{mask}}, \boldsymbol{p}^{\mathrm{mask}}(\hat{T})\right)\right], \tag{11}$$

where $\boldsymbol{y}^{\mathrm{mask}}$ is the ground truth one-hot vectors of the masked tokens. D is the training dataset. CE stands for cross-entropy loss.

As the lexicon space is high-dimensional (*e.g.* 30522-d), to avoid semantically relevant dimensions being overridden by massive non-zero values on those semantically irrelevant dimensions, we introduce a FLOPs loss (Paria et al., 2020) to encourage the sparsity of the lexicon representations:

$$\mathcal{L}_{\mathrm{FLOPs}} = \sum_{k}^{|\mathbb{W}|}(\frac{1}{B}\sum_{i=1}^{B}v_i^k)^2 + \sum_{k}^{|\mathbb{W}|}(\frac{1}{B}\sum_{i=1}^{B}t_i^k)^2, \tag{12}$$

where v_i^k and t_i^k are the activation values of the k-th dimension of the lexicon space.

The training objectives for learning lexicon representations in stage 1 and 2 are as follows:

$$\mathcal{L}^{lex} = \begin{cases}\mathcal{L}_{\mathrm{VTC}}^{lex} + \beta\cdot\mathcal{L}_{\mathrm{FLOPs}}, & \mathrm{stage1}\\ \mathcal{L}_{\mathrm{VTC}}^{lex} + \beta\cdot\mathcal{L}_{\mathrm{FLOPs}} + \mathcal{L}_{\mathrm{MLM}}, & \mathrm{stage2}\end{cases} \tag{13}$$

where β is the weight of the FLOPs loss.

3.3. Unified Learning of Latent and Lexicon Representations

Latent representations focus more on global content, while lexicon representations excel at capturing fine-grained semantics. To leverage the complementarity of latent and lexicon representations, we propose a unified learning scheme to facilitate their mutual learning.

Structure sharing. Firstly, we unify the learning of latent and lexicon representations from a structure sharing perspective. To combine the strengths of both representations, an intuitive way is to train two individual dual-encoder models for latent and lexicon representations respectively, and apply score-level fusion for retrieval. However, this parallel architecture has two drawbacks. **1)** The number of parameters and computation cost are doubled. **2)** Lacking interaction prevents them from mutual learning, and thus can't achieve the optimal performance of unified retrieval.

As shown in Figure 2, we instead propose a unified architecture, where the latent and lexicon branches share the same stem video (or text) encoder. As the two types of representations focus on information of different granularities, sharing shallow layers promotes knowledge sharing and transfer between them during the learning process. On the other hand, if the whole video (or text) encoder is shared, the raw features input to the latent and lexicon projection heads will be identical, which inevitably harms the complementarity between the two representation types. Therefore, in each stream, we introduce two representation-specific encoders on top of the stem encoder. The latent and lexicon encoders are optimized by different training objectives, and thus can learn visual and textual information of different granularities.

Self-distillation. Secondly, we use self-distillation to facilitate knowledge transfer from latent to lexicon representations. On the one hand, learning lexicon representations are more challenging due to the semantic constraints. On the other hand, while freezing the lexicon text branch in stage 1 (Section 3.2) avoids the textual lexicon distributions being corrupted, it also to some extent limits the ability of lexicon representations. As they have different focuses, the knowledge from latent representations can provide extra supervision for lexicon representation learning. Specifically, for each type of representations, we obtain the similarity scores between videos and texts by computing the dot product of their normalized representations. Denote video-to-text and text-to-video similarity scores as S_{v2t} and S_{t2v}. We use the similarity scores of latent representations as soft labels to perform self-distillation on lexicon representations, and optimize a KL divergence (D_{KL}) loss as follows:

$$\mathcal{L}_{\mathrm{D}} = D_{\mathrm{KL}}(S^{lex}_{v2t} || S^{lat}_{v2t}) + D_{\mathrm{KL}}(S^{lex}_{t2v} || S^{lat}_{t2v}). \quad (14)$$

Combining the self-distillation loss and the video-text contrastive loss of latent representations, the overall training objective of our UNIFY framework can be formulated as:

$$\mathcal{L} = \mathcal{L}^{lex} + \mathcal{L}^{lat}_{\mathrm{VTC}} + \lambda \cdot \mathcal{L}_{\mathrm{D}}, \quad (15)$$

where $\mathcal{L}^{lat}_{\mathrm{VTC}}$ has the same form as $\mathcal{L}^{lex}_{\mathrm{VTC}}$ in Equation 10, and λ is the weight of the self-distillation loss. In practice, we linearly decrease λ during training, which avoids harming the complementarity between latent and lexicon representations while facilitating lexicon representation learning.

4. Experiments

4.1. Experimental Setup

Pre-training Datasets. For a fair comparison, we follow Frozen (Bain et al., 2021) and MILES (Ge et al., 2022b) to adopt two pre-training datasets - Google Conceptual Captions (CC3M) (Sharma et al., 2018) with 3M image-text pairs, and WebVid-2M (Bain et al., 2021) with 2.5M video-text pairs.

Downstream Tasks. 1) Text-to-video retrieval. We evaluate the text-to-video retrieval performance of our UNIFY model on four mainstream datasets, *i.e.*, MSR-VTT (Xu et al., 2016), DiDeMo (Anne Hendricks et al., 2017), LSMDC (Rohrbach et al., 2015) and MSVD (Chen and Dolan, 2011). The evaluation adopts both zero-shot and fine-tuning setups, and uses Recall and Median Rank as metrics. **2) Action Recognition.** We also evaluate our model's performance in zero-shot action recognition, which can be regarded as video-to-text retrieval by describing videos with corresponding action classes following (Radford et al., 2021). Evaluation is conducted on the HMDB51 (Kuehne et al., 2011) and UCF101 (Soomro et al., 2012) datasets. Both datasets are divided into three training/test splits.

Implementation Detail. We instantiate the video encoder with the Timesformer (Bertasius et al., 2021) model, and initialize the parameters of spatial attention blocks by reusing ViT-B/16 weights pre-trained on ImageNet-21K (Ridnik et al., 2021). As for the text encoder, we use the BERT$_{base}$ (Devlin et al., 2018) model for initialization, and take its word embedding vocabulary as our lexicon. In default, the number N of shared stem blocks is set as 9, and M of representation-specific blocks is set as 3. We use 8 NVIDIA A100 GPUs for pre-training and 8 NVIDIA V100 GPUs for fine-tuning. We pre-trained UNIFY for a total of 10 epochs, during which the text encoder was frozen for the

Method	Year	Video Input	Pre-train Dataset	Pairs	MSR-VTT			
					R@1↑	R@5↑	R@10↑	MedR↓
Zero-Shot								
SupportSet (Patrick et al., 2020)	2021	R(2+1)D-34	HowTo100M	120M	12.7	27.5	36.2	24.0
Frozen (Bain et al., 2021)	2021	Raw Videos	CC3M, WebVid-2M	5.5M	18.7	39.5	51.6	10.0
AVLnet (Rouditchenko et al., 2020)	2021	ResNeXt-101	HowTo100M	120M	19.6	40.8	50.7	9.0
RegionLearner (Yan et al., 2023)	2023	Raw Videos	CC3M, WebVid-2M	5.5M	22.2	43.3	52.9	8.0
LaT (Bai et al., 2022)	2022	Raw Videos	CC3M, WebVid-2M	5.5M	23.4	44.1	53.3	8.0
MILES (Ge et al., 2022b)	2022	Raw Videos	CC3M, WebVid-2M	5.5M	26.1	47.2	56.9	7.0
BridgeFormer (Ge et al., 2022a)	2022	Raw Videos	CC3M, WebVid-2M	5.5M	26.0	46.4	56.4	7.0
TCP (Zhang et al., 2023)	2023	Raw Videos	CC3M, WebVid-2M	5.5M	26.8	48.3	57.6	7.0
UNIFY-Latent	2023	Raw Videos	CC3M, WebVid-2M	5.5M	28.4	49.4	59.4	6.0
UNIFY-Lexicon	2023	Raw Videos	CC3M, WebVid-2M	5.5M	28.0	50.2	59.8	5.0
UNIFY	2023	Raw Videos	CC3M, WebVid-2M	5.5M	**29.0**	**51.7**	**60.1**	**5.0**
Fine-Tuning								
SupportSet (Patrick et al., 2020)	2021	R(2+1)D-34	HowTo100M	120M	30.1	58.5	69.3	3.0
VideoCLIP (Xu et al., 2021b)	2021	S3D	HowTo100M	110M	30.9	55.4	66.8	-
Frozen (Bain et al., 2021)	2021	Raw Videos	CC3M, WebVid-2M	5.5M	31.0	59.5	70.5	3.0
ALPRO (Li et al., 2022b)	2022	Raw Videos	CC3M, WebVid-2M	5.5M	33.9	60.7	73.2	3.0
LaT (Bai et al., 2022)	2022	Raw Videos	CC3M, WebVid-2M	5.5M	35.3	61.3	72.9	3.0
OA-Trans (Wang et al., 2022)	2022	Raw Videos	CC3M, WebVid-2M	5.5M	35.8	63.4	76.5	3.0
RegionLearner (Yan et al., 2023)	2023	Raw Videos	CC3M, WebVid-2M	5.5M	36.3	63.9	72.5	3.0
MILES (Ge et al., 2022b)	2022	Raw Videos	CC3M, WebVid-2M	5.5M	37.7	63.6	73.8	3.0
BridgeFormer (Ge et al., 2022a)	2022	Raw Videos	CC3M, WebVid-2M	5.5M	37.6	64.8	75.1	3.0
TCP (Zhang et al., 2023)	2023	Raw Videos	CC3M, WebVid-2M	5.5M	38.0	65.5	76.4	3.0
UNIFY-Latent	2023	Raw Videos	CC3M, WebVid-2M	5.5M	40.1	66.3	75.0	2.0
UNIFY-Lexicon	2023	Raw Videos	CC3M, WebVid-2M	5.5M	40.8	68.9	78.2	2.0
UNIFY	2023	Raw Videos	CC3M, WebVid-2M	5.5M	**42.8**	**68.8**	**78.8**	**2.0**

Table 1: Text-to-video retrieval results on MSR-VTT test set. "Video Input" lists the input to the video encoder, where "Raw Videos" means training on raw video frames without pre-extracted features.

first 3 epochs. We utilize the AdamW (Loshchilov and Hutter, 2019) optimizer with a weight decay of 0.05 and batch size of 512. The learning rate was initially raised to 1e-4 in the first epoch and then decayed based on a cosine schedule. We randomly select 4 frames per video and resize to 224x224. We set the mask ratio of MLM task as 15%, and empirically set the weight of the FLOPs loss as 1e-4. As for the weight λ of self-distillation loss, we linearly decrease it from 1 to 0.

4.2. Main Results

Evaluation on Video-Text Retrieval. Table 1 shows the performance on the MSR-VTT dataset under zero-shot and fine-tuning settings. UNIFY-Latent and UNIFY-Lexicon are the retrieval results of the two representation types in our model, and UNIFY denotes the score-level fusion results of them. Based on the retrieval results, we make the following observations: 1) UNIFY-Lexicon surpasses UNIFY-Latent in almost all metrics, validating that our proposed two-stage semantics grounding approach can capture fine-grained semantic concepts and boost retrieval performance effectively. 2) UNIFY-Latent also significantly outperforms existing latent-representation-based methods, showing the unified learning scheme allows latent and lexicon representations to benefit from each other and improves performance. 3) The combination of latent and lexicon representations largely improves performance, demonstrating our UNIFY can combine the strengths of both representation types effectively. Overall, under the fine-tuning setting, UNIFY significantly outperforms TCP (Zhang et al., 2023) by 4.8%.

Table 2 shows the retrieval performance on the DiDeMo, LSMDC and MSVD datasets. Similar to MSR-VTT, our UNIFY model surpasses previous methods in almost all metrics. Notably, under the fine-tuning setting, we outperform BridgeFormer (Ge et al., 2022a) by 8.2% on the challenging DiDeMo dataset, which consists of longer videos and more complex semantic concepts compared to other datasets. The result validates the effectiveness of our approach. Another point worth noting is that on these three datasets, we list the performance of CLIP-based methods such as CLIP4Clip (Luo et al., 2022), CenterCLIP (Zhao et al., 2022) and DiCoSA (Jin et al., 2023). These models are initialized using the parameters of the CLIP (Radford et al., 2021) model, which is pre-trained on 400M image-text pair data (70$\times$ more than our pre-training data). Our method even surpasses them or achieves comparable performance, which further demonstrates the superiority of our method.

Evaluation on Action Recognition. Table 3 presents the zero-shot action recognition perfor-

Method	DiDeMo				LSMDC				MSVD			
	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
Zero-Shot												
Frozen (Bain et al., 2021)	21.1	46.0	56.2	7.0	9.3	22.0	30.1	51.0	33.7	64.7	76.3	3.0
LaT (Bai et al., 2022)	22.6	45.9	58.9	7.0	-	-	-	-	36.9	68.6	81.0	2.0
OA-Trans (Wang et al., 2022)	23.5	50.4	59.8	6.0	-	-	-	-	39.1	68.4	80.3	2.0
MILES (Ge et al., 2022b)	27.2	50.3	63.6	5.0	11.1	24.7	30.6	50.7	44.4	76.2	87.0	2.0
BridgeFormer (Ge et al., 2022a)	25.6	50.6	61.1	5.0	12.2	25.9	32.2	42.0	43.6	74.9	84.9	2.0
UNIFY-Latent	27.3	53.8	63.5	4.0	12.4	28.4	36.1	30.5	46.4	77.0	87.0	2.0
UNIFY-Lexicon	28.3	53.4	63.8	4.0	12.6	27.2	36.6	28.0	48.3	77.2	86.7	2.0
UNIFY	**29.6**	**55.5**	**66.0**	**4.0**	**14.1**	**30.4**	**37.5**	**25.0**	**48.1**	**79.7**	**87.2**	**2.0**
Fine-Tuning												
CipBert (Lei et al., 2021)	20.4	48.0	60.8	6.0	-	-	-	-	-	-	-	-
RegionLearner (Yan et al., 2023)	32.5	60.8	72.3	3.0	17.1	32.5	41.5	18.0	44.0	74.9	84.3	2.0
LaT (Bai et al., 2022)	32.6	61.3	71.6	3.0	-	-	-	-	40.0	74.6	84.2	2.0
OA-Trans (Wang et al., 2022)	34.8	64.4	75.1	3.0	18.2	34.3	43.7	18.5	-	-	-	-
ALPRO (Li et al., 2022b)	35.9	67.5	78.8	3.0	-	-	-	-	-	-	-	-
MILES (Ge et al., 2022b)	36.6	63.9	74.0	3.0	17.8	35.6	44.1	15.5	53.9	83.5	90.2	1.0
BridgeFormer (Ge et al., 2022a)	37.0	62.2	73.9	3.0	17.9	35.4	44.5	15.0	52.0	82.8	90.0	1.0
CLIP4Clip (Luo et al., 2022)	43.4	70.2	80.6	2.0	22.6	41.0	49.1	11.0	46.2	76.1	84.6	2.0
CenterCLIP (Zhao et al., 2022)	-	-	-	-	21.9	41.1	50.7	10.0	47.6	77.5	86.0	2.0
DiCoSA (Jin et al., 2023)	45.7	74.6	83.5	2.0	25.4	43.6	54.0	8.0	47.4	76.8	86.0	2.0
UNIFY-Latent	40.7	68.5	80.0	2.0	22.8	42.2	50.8	10.0	55.5	85.2	91.6	1.0
UNIFY-Lexicon	44.6	73.5	82.8	2.0	24.5	44.6	54.4	8.0	55.1	86.7	93.0	1.0
UNIFY	**45.2**	**74.0**	**83.2**	**2.0**	**24.5**	**46.3**	**55.0**	**7.0**	**57.7**	**86.8**	**92.9**	**1.0**

Table 2: Experimental results of text-to-video retrieval on the DiDeMo, LSMDC and MSVD datasets.

Method	HMDB51				UCF101			
	S1	S2	S3	Mean	S1	S2	S3	Mean
ClipBert (Lei et al., 2021)	20.0	22.0	22.3	21.4	27.5	27.0	28.8	27.8
Frozen (Bain et al., 2021)	27.5	28.3	27.7	27.8	45.4	44.7	47.7	45.9
MILES (Ge et al., 2022b)	38.4	38.6	37.8	38.3	51.8	53.4	52.8	52.7
BridgeFormer (Ge et al., 2022a)	38.0	36.1	39.1	37.7	51.1	54.3	**53.8**	53.1
UNIFY (Ours)	**38.9**	**39.6**	**39.9**	**39.5**	**53.3**	**55.0**	53.0	**53.8**

Table 3: Top-1 accuracy of zero-shot action recognition. "S" denotes different testing splits, and "Mean" is the averaged result over three splits.

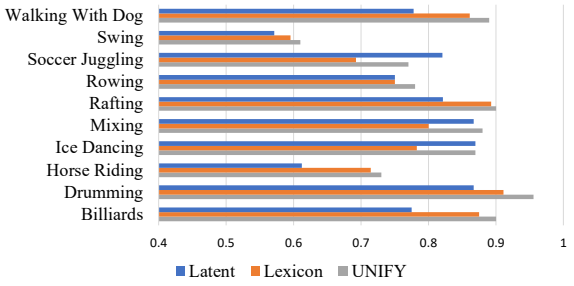

Figure 3: Zero-shot results of latent and lexicon representations on UCF101.

Figure 4: Retrieval results of two queries using latent and lexicon representations. Each row presents the top-5 ranked videos.

mance on HMDB51 and UCF101, which can be regarded as video-to-text retrieval. Except for UCF101 split 3, we outperform previous methods in all splits. This verifies that our UNIFY model can learn cross-modal representations that generalize well to the task of action recognition.

4.3. Complementarity between Latent and Lexicon Representations

Figure 3 shows the zero-shot performance of several actions in UCF101. Lexicon representations grasp fine-grained object information more effectively, showing better performance on classes like "Billiards" and "Horse Riding". On the other hand, latent representations beat lexicon representations on classes like "Mixing" and "Ice Dancing", which rely more on long-range action analysis. Moreover, combining both types of representations leads to further improvement on most classes.

Figure 4 shows the retrieval results of two queries using latent and lexicon representations respectively, where "Rank=.." indicates the ranking of the ground truth video among all candidate videos. In the first example, the lexicon representations capture the crucial detailed semantics of "pumpkin", and thus successfully retrieves the cor-

A man and a woman are performing yoga with the lady doing tricks on the guy's legs

(a) UNIFY-Lexicon -- single-stage; w/o MLM

dance	1.30	work	0.80
graceha	1.18	danced	0.80
susan	0.96	skate	0.77
hang	0.89	*	0.76
java	0.81	speechwg	0.75
yoga	1.00	blake	0.91
grace	1.00	shoes	0.89
dance	0.97	baseline	0.88
geologic	0.94	essays	0.82
bwv	0.93	act	0.82

(b) UNIFY-Lexicon -- two-stage; w/o MLM

bucks	1.30	leg	0.92
jump	1.26	dance	0.92
practice	1.14	stretching	0.89
yoga	0.96	takit	0.89
relax	0.95	strength	0.89
shoes	1.41	partner	1.03
both	1.26	positioned	0.96
yoga	1.25	tristan	0.96
legs	1.24	jump	0.94
momentum	1.18	feet	0.93

(c) UNIFY-Lexicon -- single-stage; w/MLM

jump	1.11	girl	0.98
dance	1.07	yoga	0.98
howard	1.03	auckland	0.93
jumping	1.02	feet	0.92
progress	1.01	exercise	0.91
legs	1.09	trick	0.89
yoga	1.07	woman	0.86
tricks	1.04	dance	0.86
guy	0.99	and	0.85
legrance	0.96	diveroids	0.84

(d) UNIFY-Lexicon -- two-stage; w/MLM

yoga	1.22	legs	0.90
woman	1.05	foot	0.87
exercise	1.02	performs	0.85
outside	0.94	man	0.85
jump	0.93	sport	0.82
the	1.16	doing	0.94
yoga	1.14	performing	0.89
are	1.08	feet	0.88
legs	1.04	exercises	0.88
man	1.01	on	0.85

Figure 5: Top-10 activated lexicon dimensions of four variants of UNIFY-Lexicon. Words that are semantically irrelevant to the video (or text) are highlighted in red color.

#Line	sharing stem encoder	self-distillation	UNIFY-Latent				UNIFY-Lexicon				UNIFY			
			R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
A	×	×	24.7	47.8	56.9	6.0	26.6	48.1	57.2	6.0	26.8	48.3	57.9	6.0
B	✓	×	26.9	48.2	58.7	6.0	27.5	48.5	58.9	6.0	28.2	49.5	59.3	6.0
C	✓	✓	**28.4**	**49.4**	**59.4**	**6.0**	**28.0**	**50.2**	**59.8**	**5.0**	**29.0**	**51.7**	**60.1**	**5.0**

Table 4: Ablation study on the unified learning scheme which includes structure sharing and self-distillation. Zero-shot text-to-video retrieval results on MSR-VTT are reported.

#Line	freezing strategy	MLM task	R@1↑	R@5↑	R@10↑	MedR↓
A	×	×	22.0	43.4	53.5	8.0
B	✓	×	23.2	44.3	54.2	7.0
C	×	✓	23.6	45.5	54.4	7.0
D	✓	✓	**26.6**	**48.1**	**57.2**	**6.0**

Table 5: Ablation study on the two-stage semantics grounding approach. Zero-shot text-to-video retrieval results on MSR-VTT are reported. Only the Lexicon branch is trained in this experiment.

rect video. The second query is longer and corresponds to a video with complex content. Though lexicon representations manage to capture some relevant semantics (*e.g.* man and snow), aggregating local semantics fails to grasp the overall semantics of complex texts and videos. In contrast, latent representations summarize texts and videos from a global perspective and successfully retrieves the correct video.

4.4. Two-stage Semantics Grounding Ablation

In this section, we solely train an individual lexicon branch to ablate the proposed two-stage semantics grounding (TSG) approach.

Overall TSG. Figure 5 shows the top-10 activated dimensions of different variants. In the baseline model (a) without TSG, the video falls into semantically irrelevant dimensions, and the grounding ability of the text encoder is mostly lost. Essentially, in the baseline model, lexicon representations degrade into high-dimensional latent representations. In contrast, semantically relevant dimensions are activated in model (d) which is trained using TSG. Moreover, line A and D in Table 5 show that employing TSG significantly improves the R@1 performance by 4.6%, validating the effectiveness of the proposed TSG approach.

The Freezing Strategy. Comparing (a) and (b) in Figure 5, we observe that freezing the text encoder in the first stage reduces the activation of semantically irrelevant dimensions. This demonstrates the freezing strategy prevents textual lexicon distributions from being corrupted by cross-modal alignment. Line A and B in Table 5 show that the freezing strategy improves retrieval performance.

The MLM Task. In (c) of Figure 5, although the text encoder is frozen in the first stage, the absence of the MLM task in the second stage results in the loss of semantic constraints for the text, leading to the activation of some semantically irrelevant dimensions. Comparing (c) and (d), we observe that the MLM task effectively suppresses the activation of semantically irrelevant dimensions, showing that MLM helps preserve textual semantics. Line C and D in Table 5 show combing MLM with the freezing strategy largely improves the model's performance.

4.5. Unified Learning Scheme Ablation

Structure sharing. Comparing line A and line B in Table 4, we can find that sharing the stem encoder between the latent and lexicon branches leads to an improvement in performance for both branches, lifting 1.8% and 1.7% respectively. This highlights that sharing shallow layers between the two representation types can facilitate knowledge transfer and improve the representation ability of both branches. Moreover, performance boost is also observed in the fused UNIFY results, lifting from 26.8% to 28.2% in R@1.

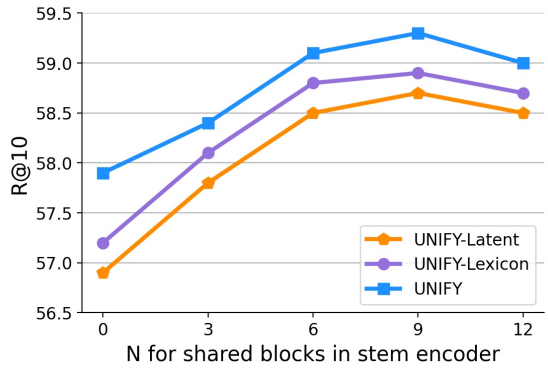

Figure 6: Effect of the number N of shared blocks in the stem encoder. Zero-shot text-to-video retrieval results on MSR-VTT are reported. Here the experiments are conducted without self-distillation.

The number of shared stem blocks. As shown in Figure 6, as we gradually increase the number of shared blocks, the performance of both the latent and lexicon branches has improved. The results demonstrate that they benefit from knowledge transfer via structure sharing. And the optimal performance is obtained at $N = 9$. However, when N continues to increase to 12, a performance decline occurs. This is because at this time the two branches completely share the encoder, and only the projection heads are different. Thus the model cannot fully learn the two specific representations. This result further verifies the rationality of the proposed structure sharing strategy.

Self-distillation. When additionally introducing the self-distillation strategy from Line B to Line C, we observe further improvement in the performance of both branches and the overall UNIFY results. Latent representations provide extra supervision information for lexicon representations, and the enhanced lexicon representations in turn inject fine-grained semantic knowledge into the stem encoder. The experimental results demonstrate the proposed self-distillation strategy can facilitate mutual learning between the two representations types and improves the performance of both UNIFY-Latent and UNIFY-Lexicon.

5. Conclusion

In this work, we presented a novel UNIFY framework, which learns lexicon representations for fine-grained semantics capturing, and unifies latent and lexicon representations for cross-modal retrieval. We proposed a two-stage semantics grounding approach to enable lexicon representations to reflect fine-grained semantic concepts. As latent and lexicon representations have different focuses, we further proposed a unified learning scheme to leverage this complementarity. Our method largely outperforms existing video-text retrieval methods, validating the effectiveness of lexicon representations and the unified learning scheme.

6. Acknowledgements

This work is supported by the National Key Research and Development Program of China (Grant No. 2022ZD0118501), Beijing Natural Science Foundation (Grant No. JQ21017, L223003, 4224093), the Natural Science Foundation of China (Grant No. 61972397, 62036011, 62192782, U2033210, 62202470), The Project of Beijing Science and technology Committee (Project No. Z231100005923046).

7. Bibliographical References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Jinbin Bai, Chunhui Liu, Feiyue Ni, Haofan Wang, Mengying Hu, Xiaofeng Guo, and Lele Cheng. 2022. Lat: latent translation with cycle-consistency for video-text retrieval. *arXiv preprint arXiv:2207.04858*.

Yang Bai, Xiaoguang Li, Gang Wang, Chaoliang Zhang, Lifeng Shang, Jun Xu, Zhaowei Wang, Fangshan Wang, and Qun Liu. 2020. Sparterm: Learning term-based sparse representation for fast text retrieval. *ArXiv*, abs/2010.00768.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, pages 813–824. PMLR.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Thibault Formal, C. Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021a. Splade v2: Sparse lexical and expansion model for information retrieval. *ArXiv*, abs/2109.10086.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. Splade: Sparse lexical and expansion model for first stage ranking. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. 2022a. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176.

Yuying Ge, Yixiao Ge, Xihui Liu, Alex Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Ping Luo. 2022b. Miles: Visual bert pre-training with injected language semantics for video-text retrieval. In *European Conference on Computer Vision*.

Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. 2023. Text-video retrieval with disentangled conceptualization and set-to-set alignment. *arXiv preprint arXiv:2305.12218*.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE.

Carlos Lassance and Stéphane Clinchant. 2022. An efficiency study for SPLADE models. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2220–2226. ACM.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, et al. 2022a. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2022b. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

Biswajit Paria, Chih-Kuan Yeh, Ian En-Hsu Yen, Ning Xu, Pradeep Ravikumar, and Barnabás Póczos. 2020. Minimizing flops to learn efficient sparse representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. 2020. Support-set bottlenecks for video-text representation learning. In *ICLR*.

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

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar

Panda, Rogerio Feris, et al. 2020. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Kai Zhang, and Daxin Jiang. 2022. Unifier: A unified retriever for large-scale retrieval. *ArXiv*, abs/2205.11194.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Alex Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2022. Object-aware video-language pre-training for retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3303–3312. IEEE.

Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021a. E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 503–513.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*.

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

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Rui Yan, Mike Zheng Shou, Yixiao Ge, Jinpeng Wang, Xudong Lin, Guanyu Cai, and Jinhui Tang. 2023. Video-text pre-training with learned regions for retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3100–3108.

Heng Zhang, Daqing Liu, Zezhong Lv, Bing Su, and Dacheng Tao. 2023. Exploring temporal concurrency for video-language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15568–15578.

Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 970–981.