

# Video-Text Retrieval by Supervised Sparse Multi-Grained Learning

**Yimu Wang**  
University of Waterloo  
yimu.wang@uwaterloo.ca

**Peng Shi**  
University of Waterloo  
peng.shi@uwaterloo.ca

## Abstract

While recent progress in video-text retrieval has been advanced by the exploration of better representation learning, in this paper, we present a novel multi-grained sparse learning framework, S3MA, to learn an aligned sparse space shared between the video and the text for video-text retrieval. The shared sparse space is initialized with a finite number of sparse concepts, each of which refers to a number of words. With the text data at hand, we learn and update the shared sparse space in a supervised manner using the proposed similarity and alignment losses. Moreover, to enable multi-grained alignment, we incorporate frame representations for better modeling the video modality and calculating fine-grained and coarse-grained similarities. Benefiting from the learned shared sparse space and multi-grained similarities, extensive experiments on several video-text retrieval benchmarks demonstrate the superiority of S3MA over existing methods. Our code is available at [link](#).

## 1 Introduction

As a fundamental task in visual-language understanding (Wang et al., 2020b; Xu et al., 2021b; Park et al., 2022a; Miyawaki et al., 2022; Fang et al., 2023a,b; Kim et al., 2023; Jian and Wang, 2023), video-text retrieval (VTR) (Luo et al., 2022; Gao et al., 2021b; Ma et al., 2022a; Liu et al., 2022a; Zhao et al., 2022; Gorti et al., 2022; Fang et al., 2022) has attracted interest from academia and industry. Although recent years have witnessed the rapid development of VTR with the support from powerful pretraining models (Luo et al., 2022; Gao et al., 2021b; Ma et al., 2022a; Liu et al., 2022a), improved retrieval methods (Bertasius et al., 2021; Dong et al., 2019; Jin et al., 2021), and video-language datasets construction (Xu et al., 2016), it remains challenging to precisely match video and language due to the raw data being in heterogeneous spaces with significant differences.

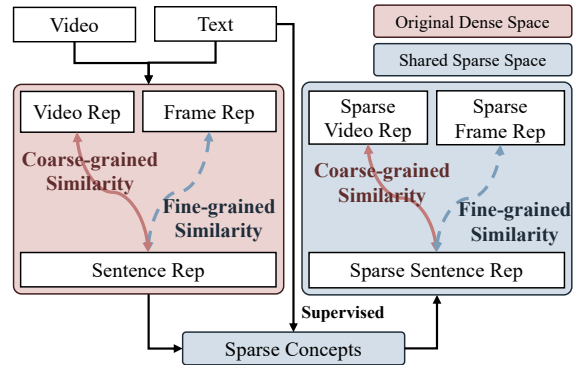


Figure 1: Our proposed *supervised shared sparse multi-grained alignment* framework for video-text retrieval maps sentence, video, and frame representations to a shared sparse space to obtain sparse sentence, video, and frame representations. Then, it calculates *coarse- and fine-grained* similarities to fully explore the power of the sparse space, which is learned in a *supervised* fashion. “Original Dense Space” represents the space containing the representations generated from modality-dependent encoders. “Shared Sparse Space” represents the space containing the sparse concepts shared across two modalities. “Rep” refers to representation.

Current VTR research (Luo et al., 2022; Ma et al., 2022a; Liu et al., 2022b) mainly aims to learn a joint feature space across modalities and then compares representations in this space. However, with the huge discrepancy between different modalities and the design of modality-independent encoders, it is challenging to directly compare and calculate the similarities between representations of different modalities generated from different encoders (Liang et al., 2022). To alleviate the mismatch caused by heterogeneous encoders and data formats, Liu et al. (2022a); Cao et al. (2022) proposed to align different modalities in a common space without supervision from text or video. However, because of the unsupervised design, the shared spaces are either randomly initialized or updated in an unsupervised fashion, which blocks the power of that aligned space. We argue that learning

a shared aligned space with supervision is a promising way to improve video-text retrieval. Borrowing from text retrieval (Karpukhin et al., 2020; Zhao et al., 2021; Gao et al., 2021a), we represent the aligned space and the space containing representations generated by modality-dependent encoders as sparse and dense spaces, respectively, as the aligned space typically carries specific semantics.

In this work, we propose a *Supervised Shared Sparse Multi-grained Alignment framework* for VTR, namely S3MA, in which the aligned sparse space is updated under the supervision of the video-text data at hand. Specifically, we initialize a finite number of sparse concepts by clustering a large number of basic concepts (words) to form the fine-grained aligned sparse space. In return, each sparse concept is composed of several words, which improves the interpretability of our model. Then, we match the sparse text and video representations effectively by projecting the video representation generated by the video encoder to this fine-grained sparse space. The sparse sentence (text) representations can be obtained by looking up the sparse concepts. To obtain sparse video representations, we first calculate the cosine similarity between the video representations and the sparse concepts. Next, by summing up all the sparse concepts with the weight of the cosine similarity between video representation and sparse concepts, we obtain the sparse video representations. Furthermore, to better match these two sparse representations, we design two loss functions to update sparse concepts, pushing the sparse representations of text and video as close as possible in the shared sparse space. This shared sparse space design not only improves the performance on VTR, but also allows us to interpret what the models have learned. The sparse aligned space, as shown in Figure 5, enables the model to accurately capture the key concepts, resulting in improved alignment within the sparse space.

Recently, Ma et al. (2022a) demonstrated that incorporating fine-grained video representations (such as frame or segment representations) with high-level video features can further improve retrieval performance. Inspired by their work, we further project *frame* representations into our designed aligned sparse space. Compared to high-level video representations, frame representations can be mapped to more detailed concepts, which enriches the overall video representations. In this way, we have fine-grained (frame) and coarse-

grained (video and sentence) representations from the sparse space and the dense space, enabling us to calculate multi-space multi-grained similarity for exploring the potential of supervised sparse space.

Finally, to evaluate the effectiveness of our proposed S3MA, we conducted experiments on three video-text benchmarks (Chen and Dolan, 2011; Fabian Caba Heilbron and Niebles, 2015; Xu et al., 2016). Benefiting from multi-grained and multi-space similarity, our proposed S3MA outperforms previous methods on all the benchmarks without requiring any additional data during training.

In summary, our contributions are as follows<sup>1</sup>:

- We propose the shared sparse space to alleviate the problem of mismatched representations from different modalities, which arises from the raw data being in heterogeneous spaces and the heterogeneous design of modality-dependent encoders.
- Our proposed S3MA achieves SOTA performance on several metrics across three VTR benchmarks.
- Detailed analysis reveals the importance of shared sparse space and multi-grained similarity. Besides, we demonstrate that the design of shared sparse space and multi-grained similarity significantly impacts retrieval performance.

## 2 Related Works

Video-Text Retrieval (VTR), which involves cross-modal alignment and abstract understanding of temporal images (videos), has been a popular and fundamental task of language-grounding problems (Wang et al., 2020a,c, 2021; Yu et al., 2023). Most existing conventional video-text retrieval frameworks (Yu et al., 2017; Dong et al., 2019; Zhu and Yang, 2020; Miech et al., 2020; Gabeur et al., 2020; Dzabraev et al., 2021; Croitoru et al., 2021) focus on learning powerful representations for video and text and extracting separated representations. Inspired by the success of self-supervised pretraining methods (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020) and vision-language pretraining (Li et al., 2020b; Gan et al., 2020; Singh et al., 2022) on large-scale unlabeled cross-modal data, recent works (Lei et al., 2021; Cheng et al., 2021; Gao et al., 2021b; Ma et al.,

<sup>1</sup>The code is released at [link](#).

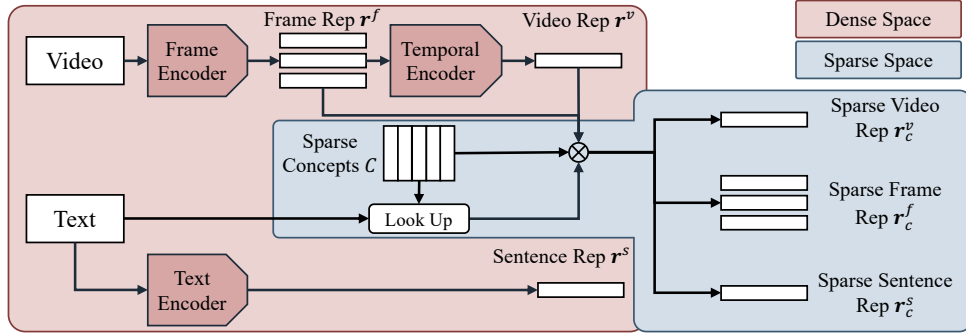


Figure 2: The illustration of representation generation in our proposed *Supervised Shared Sparse Multi-grained Alignment framework*, namely S3MA. Specifically, for multi-space alignment, we employ a shared sparse space which is consisted of a number of sparse concepts. The shared sparse space is updated in a supervised manner during the training procedure, leading to the construction of a fine-grained sparse space. “ $\otimes$ ” refers to the calculation in Eqs. (1), (2), and (3).

2022a; Park et al., 2022a; Wang et al., 2022b,c; Zhao et al., 2022; Gorti et al., 2022) have attempted to pretrain or fine-tune video-text retrieval models in an end-to-end manner. Frozen in time (Bain et al., 2021) uses end-to-end training on both image-text and video-text pairs data by uniformly sampling video frames. CLIP4Clip (Luo et al., 2022) finetunes models and investigates three similarity calculation approaches for video-sentence contrastive learning on CLIP (Radford et al., 2021). Later, to enable unsupervised sparse learning in VTR, DiscretCodebook (Liu et al., 2022a) aligns modalities in a shared space filled with concepts, which are randomly initialized and unsupervisedly updated, while VCM (Cao et al., 2022) constructs a sparse space with unsupervisedly clustered visual concepts. At the same time, OA-Trans (Wang et al., 2022a) and TABLE (Chen et al., 2023) both employ a small number of semantic tags as the input to the text encoder to improve alignment between modalities.

However, due to the unsupervised design, concepts in DiscretCodebook and VCM are either randomly initialized or updated unsupervisedly, which limits the potential of aligned sparse space. On the other hand, OA-Trans and TABLE only employ a limited number of concepts to serve as the input of the text encoder to encourage alignment. Meanwhile, these methods only perform the *coarse-grained* video-text similarity, lacking the fine-grained contrast between different modalities. In comparison, our proposed S3MA learn the aligned sparse space containing a large number of words in a *supervised* manner, under the supervision of text, and calculate frame-sentence similarity

for *multi-space multi-grained* alignment.

### 3 Methods

In this section, we introduce our proposed framework for video-text retrieval, which aligns language and video in a shared sparse space. Typically, in video-text retrieval, we have a set of examples  $\{(\mathbf{v}_i, \mathbf{t}_i)\}_{i \in [N]}$ , where  $N$  is the number of examples that are of video and language.

#### 3.1 General Video-Text Retrieval Paradigm

In this part, we present a general video-text retrieval framework widely used by previous methods (Luo et al., 2022; Liu et al., 2022a). With this paradigm, we can obtain three representations for different modalities from the dense space, *i.e.*, frame representation  $\mathbf{r}^f$ , video representation  $\mathbf{r}^v$ , and sentence representation  $\mathbf{r}^s$  by modality-dependent encoders.

**Frame and video representations:** Given a video  $\mathbf{v}$ , several video frames are first sampled as the inputs of the frame encoder to obtain the frame features  $\mathbf{r}^f \in \mathbb{R}^{n_{frame} \times d}$ , where  $n_{frame}$  is the number of frames and  $d$  is the dimension of features. As the frame representations  $\mathbf{r}^f$  are extracted through sampling, to explore the temporal correlation among different frames, we employ a temporal encoder to aggregate frame representations. With the temporal encoder and the frame representations  $\mathbf{r}^f$ , we obtain the video representations  $\mathbf{r}^v \in \mathbb{R}^{1 \times d}$ .

**Sentence representation:** Given a sentence  $\mathbf{t}$ , we use a text encoder to obtain the text representation  $\mathbf{r}^s \in \mathbb{R}^{1 \times d}$ .

### 3.2 Fine-Grained Aligned Sparse Space

The key to the video-text retrieval task is to precisely align representations from different modalities. However, due to the heterogeneous encoder architectures and data formats of different modalities, it is difficult to align directly (Liang et al., 2022). Therefore, instead of directly enforcing the representations to be aligned, we propose aligning them in an aligned sparse constructed by  $n_c$  sparse concepts  $C \in \mathbb{R}^{n_c \times d}$ . Each sparse concept  $\mathbf{c}$  represents several basic concepts (words). Moreover, to supervise the updates of sparse concepts, we utilize the human-annotated knowledge at hand, *i.e.*, text annotations in the paired video-text data.

**Initialization.** First, we map all the words into embeddings by the embedding layer  $f_{emb}$  of the text encoder. But as the number of words is relatively large (for example, in Clip (Radford et al., 2021), the number of sub-words is approximately 30k), we cluster embeddings into  $n_c$  clusters using KNN (Ganflicci, 2008) to form the sparse concepts  $C$  and represent all the words by their cluster’s centers  $\mathbf{c}$ . Consequently, each sparse concept  $\mathbf{c}$  represents a bunch of words that are similar on the embedding space, enabling fine-grained alignment. The mapping from words to sparse concepts is denoted by  $h_{w2c} \in [n_{words}] \rightarrow \{0, 1\}^{n_c \times 1}$ . Now,  $n_c$  sparse concepts have been initialized.

**Obtaining the sparse sentence representation.** For text, as the caption is at hand, we can directly tokenize the sentences into words and look up the corresponding sparse concepts in  $C$ . The sparse sentence representation  $\mathbf{r}_c^s \in \mathbb{R}^{1 \times d}$  is obtained by averaging all the representations of concepts that are fetched with the surface form of the sentence, as follows,

$$\mathbf{r}_c^s = \text{sim}^{\mathbf{t}} C / |\mathbf{t}|, \quad (1)$$

where  $|\mathbf{t}|$  is the number of words in  $\mathbf{t}$  and  $\text{sim}^{\mathbf{t}} = \sum_{w \in \mathbf{t}} h_{w2c}(w)$  is a vector with the length of  $n_c$ .

**Obtaining the sparse video representation.** We first calculate the cosine similarity  $\text{sim}^{\mathbf{v}} \in \mathbb{R}^{1 \times n_c}$  between the video representations and sparse concepts  $C$  as  $\text{sim}_j^{\mathbf{v}} = \cos(\mathbf{r}^{\mathbf{v}}, \mathbf{c}_j), \forall j \in [n_c]$ , where  $\text{sim}_j^{\mathbf{v}}$  is the  $j$ -th element of  $\text{sim}^{\mathbf{v}}$  and  $\cos(\cdot, \cdot)$  is the cosine similarity. Next, sparse video representations are obtained by weighted summing the sparse concepts as,

$$\mathbf{r}_c^{\mathbf{v}} = \text{sim}^{\mathbf{v}} C / \|\text{sim}^{\mathbf{v}}\|_1. \quad (2)$$

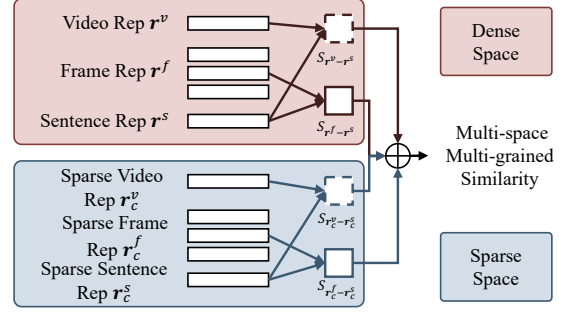


Figure 3: The illustration of similarity calculation. To enable multi-space multi-grained alignment, we calculate fine-grained (frame-sentence) and coarse-grained (video-sentence) similarity. Our preliminary experiments showed that the text encoder has a good ability to capture semantics, so we only use sentence representations for the text modality.

**Obtaining the sparse frame representation.** Similarly, the cosine similarity  $\text{sim}^{\mathbf{f}} \in \mathbb{R}^{n_{frame} \times n_c}$  between the frame representations and sparse concepts is calculated as  $\text{sim}_{i,j}^{\mathbf{f}} = \cos(\mathbf{r}_i^{\mathbf{f}}, \mathbf{c}_j), \forall i \in [n_{frame}], \forall j \in [n_c]$ , where  $\text{sim}_{i,j}^{\mathbf{f}}$  is the  $(i, j)$ -th element of  $\text{sim}^{\mathbf{f}}$  and  $\mathbf{r}_i^{\mathbf{f}}$  is the  $i$ -th row of  $\mathbf{r}^{\mathbf{f}}$ . Next, sparse frame representations are obtained as,

$$\mathbf{r}_c^{\mathbf{f}} = \sum_{i \in [n_{frame}]} \text{sim}_i^{\mathbf{f}} C / \|\text{sim}_i^{\mathbf{f}}\|_1. \quad (3)$$

Finally, we have the sparse frame, video, and sentence representations  $\mathbf{r}_c^{\mathbf{f}} \in \mathbb{R}^{n_{frame} \times d}, \mathbf{r}_c^{\mathbf{v}} \in \mathbb{R}^{1 \times d}, \mathbf{r}_c^{\mathbf{s}} \in \mathbb{R}^{1 \times d}$  with the frame and video sparse space similarity  $\text{sim}^{\mathbf{f}} \in \mathbb{R}^{n_{frame} \times n_c}$  and  $\text{sim}^{\mathbf{v}} \in \mathbb{R}^{n_c}$  along with the sentence sparse space similarity (supervision)  $\text{sim}^{\mathbf{t}}$ .

### 3.3 Multi-Space Multi-Grained Similarity

In this part, we will demonstrate our method for calculating the similarities between data from two different modalities, as shown in Figure 3, including the similarities in the dense space and in shared sparse space, inspired by Ma et al. (2022a). We can now compute multi-space (sparse and dense spaces) multi-grained (fine-grained and coarse-grained) similarity for precise alignment.

#### 3.3.1 Dense Space Similarity

**Video-Sentence similarity  $S_{\mathbf{r}^{\mathbf{v}}-\mathbf{r}^{\mathbf{s}}}$ .** To obtain a fine-grained similarity, we use a learnable matrix  $A_{\mathbf{r}^{\mathbf{v}}-\mathbf{r}^{\mathbf{s}}} \in \mathbb{R}^{d \times d}$  to focus on the discriminative features of video and sentence representations as,

$$S_{\mathbf{r}^{\mathbf{v}}-\mathbf{r}^{\mathbf{s}}} = \mathbf{r}^{\mathbf{v}} A_{\mathbf{r}^{\mathbf{v}}-\mathbf{r}^{\mathbf{s}}} \mathbf{r}^{\mathbf{s}\top}.$$



**Frame-Sentence similarity**  $S_{\mathbf{r}^f-\mathbf{r}^s}$ . To obtain a fine-grained similarity, we first calculate an *instance-aware weight* using the softmax function applied to the dot product of  $\mathbf{r}^s \mathbf{r}^f \top$ , and then use a learnable matrix  $A_{\mathbf{r}^f-\mathbf{r}^s} \in \mathbb{R}^{n_{frame} \times n_{frame}}$  to focus on discriminative frames. In this way, the similarity is calculated as,

$$S_{\mathbf{r}^f-\mathbf{r}^s} = \text{softmax}(\mathbf{r}^s \mathbf{r}^f \top) A_{\mathbf{r}^f-\mathbf{r}^s} \mathbf{r}^f \mathbf{r}^s \top.$$

### 3.3.2 Sparse Space Similarity

**Video-Sentence shared sparse space similarity**  $S_{\mathbf{r}^v-\mathbf{r}^s}$ . Similarly, to obtain a fine-grained similarity on the shared sparse space, we use a learnable matrix  $A_{\mathbf{r}^v-\mathbf{r}^s} \in \mathbb{R}^{d \times d}$  to focus on the discriminative features of sparse video and sentence representations. Now, the similarity is calculated as,

$$S_{\mathbf{r}^v-\mathbf{r}^s} = \mathbf{r}^v A_{\mathbf{r}^v-\mathbf{r}^s} \mathbf{r}^s \top.$$

**Frame-Sentence shared sparse space similarity**  $S_{\mathbf{r}_c^f-\mathbf{r}_c^s}$ . With *instance-aware weights*  $\text{softmax}(\mathbf{r}_c^s \mathbf{r}_c^f \top)$  and a learnable matrix  $A_{\mathbf{r}_c^f-\mathbf{r}_c^s} \in \mathbb{R}^{n_{frame} \times n_{frame}}$ , we get the similarity between the sparse frame and sentence representations as,

$$S_{\mathbf{r}_c^f-\mathbf{r}_c^s} = \text{softmax}(\mathbf{r}_c^s \mathbf{r}_c^f \top) A_{\mathbf{r}_c^f-\mathbf{r}_c^s} \mathbf{r}_c^f \mathbf{r}_c^s \top.$$

### 3.3.3 Overall Similarity

The overall video-text similarity is defined as,

$$S = \frac{S_{\mathbf{r}^f-\mathbf{r}^s} + S_{\mathbf{r}^v-\mathbf{r}^s} + S_{\mathbf{r}_c^f-\mathbf{r}_c^s} + S_{\mathbf{r}_c^v-\mathbf{r}_c^s}}{4}.$$

## 3.4 Objective

The objective consists of three different losses. The first component is contrastive loss. Following Clip4Clip (Luo et al., 2022), we employ the symmetric InfoNCE loss over the similarity matrix to optimize the retrieval model as,

$$\begin{aligned} \ell_{sim} &= \ell_{v2t} + \ell_{t2v} \\ &= -\frac{1}{N} \sum_{i \in [N]} \log \frac{\exp(S_{i,i})}{\sum_{j \in [N]} \exp(S_{i,j})} \\ &\quad -\frac{1}{N} \sum_{i \in [N]} \log \frac{\exp(S_{i,i})}{\sum_{j \in [N]} \exp(S_{j,i})}, \end{aligned}$$

where  $S_{i,j}$  is similarity between  $i$ -th video and  $j$ -th text and  $N$  is the number of paired data.

The second loss we minimize is the alignment loss, which matches the sparse frame and video

representations ( $\mathbf{r}_c^f$  and  $\mathbf{r}_c^v$ ) with the sparse sentence representations  $\mathbf{r}_c^s$  in the  $\ell_2$  distance, as,

$$\begin{aligned} \ell_{align} &= \frac{1}{N} \sum_{i \in [N]} \left( \|\mathbf{r}_c^v - \mathbf{r}_c^s\|_2 \right. \\ &\quad \left. + \left\| \frac{\mathbf{1} \mathbf{r}_c^f}{n_{frame}} - \mathbf{r}_c^s \right\|_2 \right), \end{aligned}$$

where  $\mathbf{1}$  is the vector only containing 1.

In addition, to match the frame and video representations with the corresponding sparse concepts, we minimize the sparse similarity loss as,

$$\begin{aligned} \ell_{sparse} &= \frac{1}{N} \sum_{i \in [N]} \left( \|sim^v - sim^t\|_2 \right. \\ &\quad \left. + \left\| \frac{\mathbf{1} sim^f}{n_{frame}} - sim^t \right\|_2 \right), \end{aligned}$$

The overall objective is the linear combination of the above three losses as,

$$\ell = \ell_{sim} + \alpha \ell_{align} + \beta \ell_{sparse},$$

where  $\alpha$  and  $\beta$  are hyperparameters controlling the trade-off between three losses. We set  $\alpha = 0.02$  and  $\beta = 0.01$  for all the experiments.

## 4 Experiments

### 4.1 Datasets and Baselines

To show the empirical efficiency of our S3MA, we train it on MSR-VTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011), and ActivityNet (Fabian Caba Heilbron and Niebles, 2015). We compare with VLM (Xu et al., 2021a), HERO (Li et al., 2020a), VideoCLIP (Xu et al., 2021b), EvO (Shvetsova et al., 2022), OA-Trans (Wang et al., 2022a), RaP (Wu et al., 2022), LiteVL (Chen et al., 2022), NCL (Park et al., 2022b), TABLE (Chen et al., 2023), VOP (Huang et al., 2023), Clip4Clip (Luo et al., 2022), X-CLIP (Ma et al., 2022a), DiscreteCodebook (Liu et al., 2022a), TS2-Net (Liu et al., 2022b), VCM (Cao et al., 2022), HiSE (Wang et al., 2022b), Align&Tell (Wang et al., 2022c), CenterCLIP (Zhao et al., 2022), and X-Pool (Gorti et al., 2022). Implementation details and evaluation protocols are deferred to the Appendix.

### 4.2 Quantitative Results

**MSR-VTT.** As shown in Table 1, S3MA achieves the best R@1 on the text-to-video retrieval results

Methods	Venue	Text-to-Video Retrieval					Video-to-Text Retrieval				
		R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
VLM	ACL'21	28.1	55.5	67.4	4.0	-	-	-	-	-	-
HERO	EMNLP'21	16.8	43.3	57.7	-	-	-	-	-	-	-
VideoCLIP	EMNLP'21	30.9	55.4	66.8	-	-	-	-	-	-	-
EvO	CVPR'22	23.7	52.1	63.7	4.0	-	-	-	-	-	-
OA-Trans	CVPR'22	35.8	63.4	76.5	3.0	-	-	-	-	-	-
RaP	EMNLP'22	40.9	67.2	76.9	2.0	-	-	-	-	-	-
<i>BLIP-based</i>											
LiteVL-S	EMNLP'22	46.7	71.8	81.7	2.0	-	-	-	-	-	-
<i>ViT-B/32-based</i>											
Align&Tell	TMM	45.2	73.0	82.9	2.0	-	43.4	70.9	81.8	2.0	-
X-Pool	CVPR'22	46.9	72.8	82.2	2.0	14.3	-	-	-	-	-
CenterCLIP	SIGIR'22	44.2	71.6	82.1	2.0	15.1	42.8	71.7	82.2	2.0	10.9
TS2-Net	ECCV'22	47.0	<u>74.5</u>	<u>83.8</u>	2.0	<u>13.0</u>	45.3	74.1	83.7	2.0	9.2
X-CLIP	ACM MM'22	46.1	74.3	83.1	2.0	13.2	46.8	73.3	84.0	2.0	<u>9.1</u>
NCL	EMNLP'22	43.9	71.2	81.5	2.0	15.5	44.9	71.8	80.7	2.0	12.8
TABLE	AAAI'23	47.1	74.3	82.9	2.0	13.4	47.2	<u>74.2</u>	<u>84.2</u>	2.0	11.0
VOP	CVPR'23	44.6	69.9	80.3	2.0	16.3	44.5	70.7	80.6	2.0	11.5
CLIP4Clip	NC	44.5	71.4	81.6	2.0	15.3	-	-	-	-	-
DiscreteCodebook	ACL'22	43.4	72.3	81.2	-	14.8	42.5	71.2	81.1	-	12.0
VCM	AAAI'22	43.8	71.0	-	2.0	14.3	45.1	72.3	82.3	2.0	10.7
S3MA		<u>49.1</u>	73.9	82.8	2.0	13.5	<u>46.9</u>	73.8	82.1	2.0	9.3
S3MA†		<b>51.7</b>	<b>75.9</b>	<b>85.4</b>	1.0	<b>11.1</b>	<b>51.6</b>	<b>76.8</b>	<b>85.0</b>	1.0	<b>8.4</b>
<i>ViT-B/16-based</i>											
Align&Tell	TMM	47.4	74.3	84.1	2.0	-	45.3	73.5	83.7	2.0	-
CenterCLIP	SIGIR'22	48.4	73.8	82.0	2.0	13.8	<u>47.7</u>	75.0	83.3	2.0	10.2
HiSE	ACM MM'22	45.0	72.7	81.3	2.0	-	<u>46.6</u>	73.3	82.3	2.0	-
TS2-Net	ECCV'22	49.4	<u>75.6</u>	<u>85.3</u>	2.0	13.5	46.6	75.9	<u>84.9</u>	2.0	<u>8.9</u>
CLIP4Clip	NC	45.8*	74.3*	84.1*	2.0*	-	43.2*	71.3*	82.0*	2.0*	-
S3MA		<u>49.8</u>	75.1	83.9	2.0	12.2	47.3	<u>76.0</u>	84.3	2.0	<u>8.9</u>
S3MA†		<b>53.1</b>	<b>78.2</b>	<b>86.2</b>	1.0	<b>10.5</b>	<b>52.7</b>	<b>79.2</b>	<b>86.3</b>	1.0	<b>8.2</b>

Table 1: Video-Text retrieval results on MSR-VTT. \* represents data copied from Align&Tell. The best results are marked in **bold**. The second best results are underlined. “NC” refers to Neurocomputing. † refers to the results with the inverted softmax.

Methods	Venue	Text-to-Video Retrieval			
		R@1↑	R@5↑	R@10↑	MnR↓
<i>MSVD</i>					
X-CLIP	ACM MM'22	47.1	<u>77.8</u>	-	<u>9.5</u>
HiSE	ACM MM'22	45.9	76.2	84.6	-
X-Pool	CVPR'22	<u>47.2</u>	77.4	<b>86.0</b>	<b>9.3</b>
CLIP4Clip	NC	45.2	75.5	84.3	10.3
S3MA		<b>47.3</b>	<b>78.8</b>	<u>85.7</u>	<b>9.3</b>
<i>ActivityNet</i>					
Align&Tell	TMM	42.6	73.8	-	-
X-CLIP	ACM MM'22	<u>44.3</u>	<u>74.1</u>	-	7.9
TS2-Net	ECCV'22	41.0	73.6	<u>84.5</u>	8.4
CLIP4Clip	NC	40.5	72.4	-	7.5
VCM	AAAI'22	40.8	72.8	-	<u>7.3</u>
S3MA		<b>45.0</b>	<b>75.5</b>	<b>85.7</b>	<b>6.3</b>

Table 2: Text-Video retrieval results on MSVD and ActivityNet. The best results are marked in **bold**. The second best results are underlined.

using ViT-B/32 and ViT-B/16, outperforming the second-best method by 2.1 and 0.4, respectively.

The performance of S3MA on the video-to-text retrieval task is also comparable with previous methods, achieving the best and second-best results on R@1 and R@5 using ViT-B/32. Moreover, we notice that only 1 previous method using ViT-B/16 outperforms S3MA with ViT-B/32 on the text-to-video retrieval, demonstrating the effectiveness of S3MA. Compared to DiscreteCodebook (Liu et al., 2022a), which aligns modalities in an unsupervised manner, S3MA outperforms DiscreteCodebook on every metric. Meanwhile, S3MA also outperforms VCM (Cao et al., 2022), which constructs an aligned space with unsupervisedly clustered visual concepts, demonstrating the importance of supervising alignment in the sparse space. This suggests that aligning modalities with fine-grained supervision is a promising approach to improving video-to-text retrieval performance.

**MSVD and ActivityNet.** The results on MSVD

	Text-to-Video Retrieval					Video-to-Text Retrieval				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
S3MA (ViT-B/32) w. SE	47.3	73.5	82.0	2.0	13.3	45.6	73.4	<b>82.4</b>	2.0	<b>9.1</b>
S3MA (ViT-B/32) w. Emb	<b>49.1</b>	<b>73.9</b>	<b>82.8</b>	2.0	<b>13.5</b>	<b>46.9</b>	<b>73.8</b>	82.1	2.0	9.3

Table 3: Comparing the power of different sparse spaces on MSR-VTT. “Emb” and “SE” refers to the embedding space and semantic embedding space.

	Text-to-Video Retrieval					Video-to-Text Retrieval				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
S3MA (ViT-B/32) w/o clustering	48.7	<b>74.4</b>	<b>83.0</b>	2.0	<b>13.4</b>	46.7	73.3	<b>82.6</b>	2.0	<b>9.2</b>
S3MA (ViT-B/32)	<b>49.1</b>	73.9	82.8	2.0	13.5	<b>46.9</b>	<b>73.8</b>	82.1	2.0	9.3

Table 4: Ablation study on the effect of clustering when constructing the shared sparse space.

Size	Text-to-Video Retrieval			Video-to-Text Retrieval		
	R@1	R@5	MnR	R@1	R@5	MnR
512	48.7	73.0	<b>12.9</b>	46.4	72.8	<b>9.0</b>
1024	<b>49.1</b>	<b>73.9</b>	<u>13.5</u>	<u>46.9</u>	<b>73.8</b>	9.3
2048	48.3	<b>73.9</b>	<u>13.5</u>	<b>47.0</b>	72.7	<u>9.1</u>
4096	47.6	<u>73.6</u>	13.6	46.8	<u>73.4</u>	9.3
DC (1024)	43.4	72.3	14.8	42.5	71.2	12.0
VCM	43.8	71.0	14.3	45.1	72.3	10.7

Table 5: Retrieval performance with different sizes of sparse space on the MSR-VTT dataset using S3MA with ViT/B-32. “DC” represents DiscreteCodebook (Liu et al., 2022a), which also aligns modalities in a sparse space whose size is 1024 with the base model of ViT/B-32. The best results are marked in **bold**. The second best results are underlined.

and ActivityNet are shown in Table 2. S3MA achieves the best R@1 on text-to-video retrieval on two datasets compared to the previous methods. Besides, with the shared sparse space and multi-grained alignment, S3MA also has the lowest MnR.

### 4.3 Ablation Studies

In this part, we present a series of ablation experiments on MSR-VTT to demonstrate the effectiveness of different components of S3MA. The evaluation of two proposed losses, similarity calculation, and the importance of word-level features are deferred to the Appendix.

#### 4.3.1 Efficiency of Sparse Space

**The choice of different initialization of sparse spaces.** To choose the best initialization method for the sparse space, we conduct experiments using two different initializations, *i.e.*, the embedding and semantic embedding spaces, as shown in Table 3. The embedding space is the one we use in S3MA,

while the semantic embedding space, is initialized by outputs of the last layer in the text encoder, with input consisting of a word and two [SEP] tokens. By replacing the embedding initialization with the semantic embedding, the retrieval performance of S3MA decreases, proving the superiority of embedding space over the semantic embedding space.

**Size of sparse space.** Another important factor to consider is the size of the sparse space. When we have unlimited data to train models, a large sparse space is ideal. However, when the data is limited, a large sparse space can lead to sparse gradients, resulting in most of the concepts not being able to be updated, while a small sparse space will restrict the retrieval ability as it becomes more challenging to distinguish between numerous data points. The results of these experiments can be found in Table 5. We see that halving and doubling the size of the sparse space slightly decreases performance.

**Impact of clustering.** As S3MA clusters all the embeddings to initialize concept clusters, it is uncertain whether clustering will hinder the power of the shared sparse space. Clustering can be useful to extract high-level abstract concepts and reduce noise. However, it may also lead to a loss of information, which is important for fine-grained alignment. Specifically, we compare the performance of S3MA to that of a modified version, S3MA w/o clustering concepts, which directly uses over 30k basic concepts to form the shared sparse space. Quantitative results can be found in Table 4. The results show that without clustering, R@5, R@10, and MnR on text-to-video retrieval and R@10 and MnR on video-to-text retrieval are improved. On one hand, similar basic concepts can be better separated, which leads to more precise alignment. On the other hand, that may lead to

Dense Space		Sparse Space		Text-to-Video Retrieval					Video-to-Text Retrieval				
S-V	S-F	S-V	S-F	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
✓				42.8	72.0	82.3	2.0	15.0	41.9	71.1	81.5	2.0	11.1
✓		✓		43.3	70.5	81.4	2.0	15.6	42.5	71.0	80.9	2.0	11.9
	✓			44.4	71.8	81.8	2.0	14.5	44.1	71.8	81.7	2.0	10.4
	✓		✓	44.8	72.1	81.7	2.0	15.9	41.7	70.2	79.6	2.0	10.8
✓			✓	42.9	72.3	81.6	2.0	15.2	42.0	70.9	81.1	2.0	11.0
	✓	✓		43.8	72.1	82.3	2.0	14.7	41.5	70.6	80.3	2.0	9.8
✓	✓			44.0	71.3	80.9	2.0	14.8	43.6	69.5	80.1	2.0	10.4
✓	✓	✓		47.4	73.3	82.4	2.0	<b>12.9</b>	46.4	73.0	<b>82.2</b>	2.0	<b>8.9</b>
✓	✓		✓	47.4	73.6	82.5	2.0	13.2	<b>47.3</b>	72.3	81.7	2.0	<b>8.9</b>
✓	✓	✓	✓	<b>49.1</b>	<b>73.9</b>	<b>82.8</b>	2.0	13.5	46.9	<b>73.8</b>	82.1	2.0	9.3

Table 6: Retrieval performance with different similarities on MSR-VTT using S3MA with the base model of ViT-B/32. “S-V” and “S-F” represent Sentence-Video (coarse-grained) and Sentence-Frame (fine-grained) similarities.

Base Model	TE	Text-to-Video			Video-to-Text		
		R@1	R@5	MnR	R@1	R@5	MnR
ViT-B/32		47.0	<b>73.9</b>	14.5	45.7	72.3	9.6
	✓	<b>49.1</b>	<b>73.9</b>	<b>13.5</b>	<b>46.9</b>	<b>73.8</b>	<b>9.3</b>
ViT-B/16		47.3	74.9	12.8	46.1	75.1	9.5
	✓	<b>49.8</b>	<b>75.1</b>	<b>12.2</b>	<b>47.3</b>	<b>76.0</b>	<b>8.9</b>

Table 7: Retrieval performance with or without the temporal encoder (“TE”) and with different base models.

sparse gradients, resulting in some concepts not being fully updated while others are over-updated. This might cause some concepts to be under or over-represented, which might negatively impact the performance (Radovanovic et al., 2010). Therefore, it’s important to find the balance in clustering to achieve the best performance.

### 4.3.2 Efficiency of Multi-Grained Similarities

In order to fully evaluate the impact of multi-grained similarities, we compare different variants of S3MA and the results are shown in Table 6. From these results, we can draw three conclusions,

- Multi-grained similarities are crucial for retrieval. Using both coarse- and fine-grained alignments in the dense space improved R@1 from 42.8 and 41.9 to 44.0 and 43.6 on text-to-video and video-to-text retrieval compared with only using coarse-grained alignment in the dense space, respectively. The same observation can be observed in the sparse space.
- Sparse space plays a crucial role in improving the alignment of modalities. We observe that incorporating coarse-grained in the dense and sparse spaces improves R@1 for text-to-video

retrieval from 42.8 to 43.3 compared to only performing coarse-grained similarity in the dense space, respectively.

- Using multi-space and multi-grained similarities simultaneously achieves the best performance. R@1 on text-to-video and video-to-text retrieval is significantly improved from 42.8 and 41.9 to 49.1 and 46.9, respectively.

### 4.3.3 Temporal Encoder and Larger Model

We also investigate the effect of the temporal encoder (TE, a small sequence transformer) and different base models. The results are shown in Table 7. S3MA with TE outperforms S3MA without TE, because it is able to better model the temporal relation among different frames in a video. Besides, using a larger base model, such as ViT-B/16, further improves the performance of S3MA, as a larger base model typically has better representation learning abilities benefiting this retrieval task as well. Similar conclusions can be found in previous works (Luo et al., 2022; Ma et al., 2022a).

## 4.4 Qualitative Results

To qualitatively validate the effectiveness of S3MA and the alignment in the sparse space, we present examples of video-to-text and text-to-video retrieval on MSR-VTT in Figures 4, 6 and 7, and the alignment in sparse space in Figure 5, respectively. The retrieval results show the satisfactory performance of S3MA, benefiting from multi-space multi-grained similarity. Notably, S3MA demonstrates precise identification of the color (*green*), objects (*bicycle*), and humans (*a man*), indicating its proficiency in capturing intricate details. In Fig-





Figure 4: Video-Text retrieval examples.

Query: a(519) movie(947) director(694) talking(248) to(1017) the(519) media(154) men(28) in(1017) press(915) conference(133) regarding(827) his(384) movie(947) and(522) hero(213) also(41)

Video Sparse Similarity – Top10 Indices and Similarities

Ind	519	1017	41	28	213	522	140	248	827	578
Sim	0.93	0.91	0.80	0.70	0.63	0.57	0.53	0.50	0.44	0.42

Frame Sparse Similarity – Top10 Indices and Similarities

Ind	519	213	1017	41	248	827	522	124	28	140
Sim	0.90	0.85	0.83	0.82	0.82	0.76	0.75	0.71	0.68	0.68

Figure 5: An example of alignment on the sparse space. The index of the concepts is shown in the brackets.

ure 5, we notice that, the video and frame features are perfectly aligned with the corresponding sparse concepts as exhibiting high similarities.

## 5 Conclusion

In this paper, to better align video and text modalities, we proposed a multi-space, multi-grained video-text retrieval framework, S3MA. Specifically, S3MA aligned different modalities in a fine-grained shared sparse space, which is initialized with a finite number of concept clusters consisting of a number of basic concepts (words) and updated in a supervised fashion with the guide of text. Besides, S3MA employed frame (fine-grained) and video (coarse-grained) features to encourage models to perform multi-grained similarity alignment. Finally, we conducted extensive experiments on three representative video-text retrieval benchmarks, showing the superiority of S3MA.

## Limitations

In the future, it would be promising to seek more fine-grained alignment, such as instance (object)-level or word-level alignment, for aligning different modalities. Moreover, our experiment focused solely on the application of sparse retrieval in video-text retrieval. It would be great to see whether sparse retrieval can help other cross-modal retrieval tasks, e.g., audio-text, image-text, audio-video, and audio-image retrieval. Additionally, incorporating more detailed information such as the relationship between different objects and frames would be beneficial for the video-text retrieval problem.

Regarding the sparse space, we notice that some sparse concepts are retrieved a lot during the training procedure which might lead to the emergence of hubness (Radovanovic et al., 2010). Investigating improved clustering methods to mitigate hubness would be an interesting direction for future research. That might be due to the KNN clustering strategy and in the future and introducing better clustering strategies might be able to reduce the hubness issue, such as weighted KNN, semantic-based KNN, or part-of-speech tagging-based KNN.

## References

- 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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1708–1718. IEEE.
- Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. *Is space-time attention all you need for video understanding?* In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. *Language models are few-shot learners*. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

- Shuqiang Cao, Bairui Wang, Wei Zhang, and Lin Ma. 2022. [Visual consensus modeling for video-text retrieval](#). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 167–175. AAAI Press.
- David Chen and William 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, Portland, Oregon, USA. Association for Computational Linguistics.
- Dongsheng Chen, Chaofan Tao, Lu Hou, Lifeng Shang, Xin Jiang, and Qun Liu. 2022. [LiteVL: Efficient video-language learning with enhanced spatial-temporal modeling](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7985–7997, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yizhen Chen, Jie Wang, Lijian Lin, Zhongang Qi, Jin Ma, and Ying Shan. 2023. [Tagging before Alignment: Integrating Multi-Modal Tags for Video-Text Retrieval](#). In *AAAI Conference on Artificial Intelligence*. arXiv. ArXiv:2301.12644 [cs].
- Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. 2021. [Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss](#). *CoRR*, abs/2109.04290.
- Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. 2021. [Teachtext: Crossmodal generalized distillation for text-video retrieval](#). In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11563–11573. IEEE.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [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. Association for Computational Linguistics.
- Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. [Dual encoding for zero-example video retrieval](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9346–9355. Computer Vision Foundation / IEEE.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. 2021. [MDMMT: multidomain multimodal transformer for video retrieval](#). In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3354–3363. Computer Vision Foundation / IEEE.
- Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. [Activitynet: A large-scale video benchmark for human activity understanding](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.
- Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. 2022. [Multi-modal cross-domain alignment network for video moment retrieval](#). *IEEE Transactions on Multimedia*.
- Xiang Fang, Daizong Liu, Pan Zhou, and Guoshun Nan. 2023a. [You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2448–2460.
- Xiang Fang, Daizong Liu, Pan Zhou, Zichuan Xu, and Ruixuan Li. 2023b. [Hierarchical local-global transformer for temporal sentence grounding](#). *IEEE Transactions on Multimedia*.
- Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. [Multi-modal transformer for video retrieval](#). In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 214–229. Springer.
- Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. [Large-scale adversarial training for vision-and-language representation learning](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021a. [COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3030–3042, Online. Association for Computational Linguistics.

- Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. 2021b. [CLIP2TV: an empirical study on transformer-based methods for video-text retrieval](#). *CoRR*, abs/2111.05610.
- F. Gianfeli. 2008. [Nearest-neighbor methods in learning and vision](#). *IEEE Transactions on Neural Networks*, 19(2):377–377.
- Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. 2022. [X-pool: Cross-modal language-video attention for text-video retrieval](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4996–5005. IEEE.
- Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. 2023. [VoP: Text-Video Co-Operative Prompt Tuning for Cross-Modal Retrieval](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6565–6574.
- Xiangru Jian and Yimu Wang. 2023. [Invgc: Robust cross-modal retrieval by inverse graph convolution](#).
- Weike Jin, Zhou Zhao, Pengcheng Zhang, Jieming Zhu, Xiuqiang He, and Yueting Zhuang. 2021. [Hierarchical cross-modal graph consistency learning for video-text retrieval](#). In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1114–1124. ACM.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Taehoon Kim, Pyunghwan Ahn, Sangyun Kim, Sihaeng Lee, Mark Marsden, Alessandra Sala, Seung Hwan Kim, Bohyung Han, Kyoung Mu Lee, Honglak Lee, Kyoungsoon Bae, Xiangyu Wu, Yi Gao, Hailiang Zhang, Yang Yang, Weili Guo, Jianfeng Lu, Youngtaek Oh, Jae Won Cho, Dong jin Kim, In So Kweon, Junmo Kim, Wooyoung Kang, Won Young Jho, Byungseok Roh, Jonghwan Mun, Solgil Oh, Kenan Emir Ak, Gwang-Gook Lee, Yan Xu, Mingwei Shen, Kyomin Hwang, Wonsik Shin, Kamin Lee, Wonhark Park, Dongkwan Lee, Nojun Kwak, Yujin Wang, Yimu Wang, Tiancheng Gu, Xingchang Lv, and Mingmao Sun. 2023. [Nice: Cvpr 2023 challenge on zero-shot image captioning](#).
- 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7331–7341. Computer Vision Foundation / IEEE.
- Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020a. [HERO: Hierarchical encoder for Video+Language omni-representation pre-training](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online. Association for Computational Linguistics.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. [Oscar: Object-semantics aligned pre-training for vision-language tasks](#). In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer.
- Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. [Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning](#). In *Advances in neural information processing systems*.
- Alexander Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2022a. [Cross-modal discrete representation learning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3013–3035, Dublin, Ireland. Association for Computational Linguistics.
- Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. 2022b. [Ts2-net: Token shift and selection transformer for text-video retrieval](#). In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, volume 13674 of *Lecture Notes in Computer Science*, pages 319–335. Springer.
- Ilya Loshchilov and Frank Hutter. 2017. [SGDR: stochastic gradient descent with warm restarts](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- 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.
- Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022a. [X-CLIP: end-to-end multi-grained contrastive learning for video-text retrieval](#). In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 638–647. ACM.



- Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022b. [X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval](#). In *ACM international conference on multimedia*, MM '22, pages 638–647, New York, NY, USA. Association for Computing Machinery. Number of pages: 10 Place: Lisboa, Portugal.
- 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9876–9886. Computer Vision Foundation / IEEE.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. [Howto100m: Learning a text-video embedding by watching hundred million narrated video clips](#). In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.
- Shumpei Miyawaki, Taku Hasegawa, Kyosuke Nishida, Takuma Kato, and Jun Suzuki. 2022. [Scene-text aware image and text retrieval with dual-encoder](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–433, Dublin, Ireland. Association for Computational Linguistics.
- Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. 2022a. [Exposing the limits of video-text models through contrast sets](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586, Seattle, United States. Association for Computational Linguistics.
- Yookoon Park, Mahmoud Azab, Seungwhan Moon, Bo Xiong, Florian Metz, Gourab Kundu, and Kirmani Ahmed. 2022b. [Normalized contrastive learning for text-video retrieval](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 248–260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. [Hubs in space: Popular nearest neighbors in high-dimensional data](#). *J. Mach. Learn. Res.*, 11:2487–2531.
- Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio Feris, David Harwath, James Glass, and Hilde Kuehne. 2022. [Everything at Once – Multi-modal Fusion Transformer for Video Retrieval](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19988–19997, New Orleans, LA, USA. IEEE.
- Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. [FLAVA: A foundational language and vision alignment model](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15617–15629. IEEE.
- Alex Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2022a. [Object-aware Video-language Pre-training for Retrieval](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3303–3312, New Orleans, LA, USA. IEEE.
- Haoran Wang, Di Xu, Dongliang He, Fu Li, Zhong Ji, Jungong Han, and Errui Ding. 2022b. [Boosting video-text retrieval with explicit high-level semantics](#). In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4887–4898. ACM.
- Xiaohan Wang, Linchao Zhu, Zhedong Zheng, Mingliang Xu, and Yi Yang. 2022c. [Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision](#). *IEEE Transactions on Multimedia*, pages 1–11.
- Yimu Wang, Shiyin Lu, and Lijun Zhang. 2020a. [Searching privately by imperceptible lying: A novel private hashing method with differential privacy](#). In *Proceedings of the 28th ACM International Conference on Multimedia*, page 2700–2709.
- Yimu Wang, Ren-Jie Song, Xiu-Shen Wei, and Lijun Zhang. 2020b. [An adversarial domain adaptation network for cross-domain fine-grained recognition](#). In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1217–1225.
- Yimu Wang, Xiu-Shen Wei, Bo Xue, and Lijun Zhang. 2020c. [Piecewise hashing: A deep hashing method for large-scale fine-grained search](#). In *Pattern Recognition and Computer Vision - Third Chinese Conference, PRCV 2020, Nanjing, China, October 16-18, 2020, Proceedings, Part II*, pages 432–444.
- Yimu Wang, Bo Xue, Quan Cheng, Yuhui Chen, and Lijun Zhang. 2021. [Deep unified cross-modality hashing by pairwise data alignment](#). In *Proceedings*



- of the Thirtieth International Joint Conference on Artificial Intelligence, *IJCAI-21*, pages 1129–1135.
- Xing Wu, Chaochen Gao, Zijia Lin, Zhongyuan Wang, Jizhong Han, and Songlin Hu. 2022. **RaP: Redundancy-aware video-language pre-training for text-video retrieval**. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3036–3047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Mousouh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021a. **VLM: Task-agnostic video-language model pre-training for video understanding**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239, Online. Association for Computational Linguistics.
- Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021b. **Video-CLIP: Contrastive pre-training for zero-shot video-text understanding**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. **MSR-VTT: A large video description dataset for bridging video and language**. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society.
- Qiyang Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. 2023. **Multimodal federated learning via contrastive representation ensemble**. In *The Eleventh International Conference on Learning Representations*.
- Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. **A joint sequence fusion model for video question answering and retrieval**. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 487–503. Springer.
- Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. **End-to-end concept word detection for video captioning, retrieval, and question answering**. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3261–3269. IEEE Computer Society.
- 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, SIGIR '22*, page 970–981, New York, NY, USA. Association for Computing Machinery.
- Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. **SPARTA: Efficient Open-Domain Question Answering via Sparse Transformer Matching Retrieval**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 565–575, Online. Association for Computational Linguistics.
- Linchao Zhu and Yi Yang. 2020. **Actbert: Learning global-local video-text representations**. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8743–8752. Computer Vision Foundation / IEEE.

## A Experiments

### A.1 Datasets Details

**MSR-VTT** (Xu et al., 2016) contains 10,000 videos with length varying from 10 to 32 seconds, each paired with about 20 human-labeled captions. Following the evaluation protocol from previous works Yu et al. (2018); Miech et al. (2019), we use the training-9k / test 1k-A splits for training and testing respectively.

**MSVD** (Chen and Dolan, 2011) contains 1,970 videos with a split of 1200, 100, and 670 as the train, validation, and test set, respectively. The duration of videos varies from 1 to 62 seconds. Each video is paired with 40 English captions.

**ActivityNet** (Fabian Caba Heilbron and Niebles, 2015) is consisted of 20,000 Youtube videos with 100,000 densely annotated descriptions. For a fair comparison, following the previous setting (Luo et al., 2022; Gabeur et al., 2020), we concatenate all captions together as a paragraph to perform a video-paragraph retrieval task by concatenating all the descriptions of a video. Performances are reported on the “val1” split of the ActivityNet.

### A.2 Implementation Details and Evaluation Protocols

Following Luo et al. (2022); Ma et al. (2022a), we use a standard vision transformer (Dosovitskiy et al., 2021) with 12 layers which are initialized with the public CLIP (Radford et al., 2021) checkpoints. We directly use the text encoder of CLIP as our text encoder which is also initialized with the public CLIP checkpoints.

We set the query, key, and value projection dimension size as 512 to match CLIP’s output dimension and we initialize our logit scaling parameter  $\lambda$  with the value from the pre-trained CLIP model. All models are optimized for 5 epochs on MSR-VTT and MSVD, and for ActivityNet, the models are trained for 20 epochs. We use AdamW (Loshchilov and Hutter, 2019) with a weight decay of 0.2 and decay the learning rate using a cosine schedule (Loshchilov and Hutter, 2017), following the method used in CLIP (Radford et al., 2021). For all experiments, we uniformly sample 12 frames from every video, resizing each frame to 224x224 as per previous works (Luo et al., 2022; Ma et al., 2022a). we set  $n_{codes} = 1024$  following DiscreteCodebook (Liu et al., 2022a). To evaluate the retrieval performance of our proposed model, we use recall at Rank K

(R@K, higher is better), median rank (MdR, lower is better), and mean rank (MnR, lower is better) as retrieval metrics, which are widely used in previous retrieval works (Radford et al., 2021; Luo et al., 2022; Ma et al., 2022a).

### A.3 Ablation Studies

**Evaluating the calculation of similarity between video and frame representations and cluster concepts in S3MA.** In S3MA, we use cosine similarity to calculate  $sim^f$  and  $sim^v$ . Another way of calculating  $sim^f$  and  $sim^v$  might be using multi-label classification. To compare the effect of multi-label classification and cosine similarity, we conduct experiments using two multi-layer perceptrons (MLPs) with two layers and the *ReLU* activation to predict the similarity between video and frame representations and cluster concepts. Two MLPs are also trainable. Quantitative results are shown in Table 8. Our quantitative results, shown in Table 8, indicate that the use of MLPs decreases R@1 on text-to-video and video-to-text retrieval. This suggests that cosine similarity is more suitable for VTR.

**Evaluating the importance of supervised alignment in S3MA.** In S3MA, the aligned sentence representation  $\mathbf{r}_c^s$  is obtained from the text as in Eq. (1). This process aligns the sentence representation based on the instruction of the text. By doing so, the aligned sentence representation  $\mathbf{r}_c^s$  can serve as the supervision (an anchor) for aligning video and frame features, providing a reference point for the alignment of different modalities. To investigate the importance of placing an anchor  $\mathbf{r}_c^s$  for better alignment, we compare it to obtaining aligned sentence representation through the similarity between concept clusters  $C$  and sentence feature  $\mathbf{r}^t$ . This alternative approach allows us to evaluate the effectiveness of using an anchor for alignment and to understand how it improves the performance of the model. To investigate the alternative approach of obtaining aligned sentence representation without an anchor, we calculate the sentence sparse space similarity  $sim^t \in \mathbb{R}^{1 \times n_c}$  by calculating the cosine similarity between sentence representations and concepts as  $sim_j^t = \cos(\mathbf{r}^s, C_j)$ , where  $sim_j^t$  is the  $j$ -th element of  $sim^t$ ,  $C_j$  is the  $j$ -th row of  $C$ , and  $\cos$  is the cosine similarity. The aligned sentence representation  $\mathbf{r}^t$  without the instruction of text is obtained by matrix multiplication as follows:

$$\mathbf{r}^t = sim^t C / \|sim^t\|_1, \quad (4)$$

	Text-to-Video Retrieval					Video-to-Text Retrieval				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
S3MA (ViT-B/32) w. multi-label classification	47.0	73.6	<b>82.9</b>	2.0	<b>12.5</b>	45.5	<b>73.8</b>	<b>82.8</b>	2.0	<b>8.7</b>
S3MA (ViT-B/32) w. cosine	<b>49.1</b>	<b>73.9</b>	82.8	2.0	13.5	<b>46.9</b>	<b>73.8</b>	82.1	2.0	9.3

Table 8: Ablation study on the calculation of similarity between video and frame representations and cluster concepts.

	Text-to-Video Retrieval					Video-to-Text Retrieval				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
S3MA (ViT-B/32) w/o anchor	47.8	72.9	82.3	2.0	<b>13.4</b>	46.4	<b>74.9</b>	<b>82.1</b>	2.0	<b>9.1</b>
S3MA (ViT-B/32) w. anchor	<b>49.1</b>	<b>73.9</b>	<b>82.8</b>	2.0	13.5	<b>46.9</b>	73.8	<b>82.1</b>	2.0	9.3

Table 9: Ablation study on the instruction of text, *i.e.*, generating  $r_c^s$  using the similarity or the text. “w. anchor” refers to obtain  $r_c^s$  by text as Eq. (1). “w/o anchor” refers to obtain  $r_c^s$  by the similarity between sentence representations and concepts  $C$  as Eq. (4)

$\ell_{align}$	$\ell_{alignsim}$	Text-to-Video Retrieval					Video-to-Text Retrieval				
		R@1↑	R@5↑	R@10↑	MnR↓	MeanR↓	R@1↑	R@5↑	R@10↑	MnR↓	MeanR↓
✓	✓	48.0	72.9	82.4	2.0	13.5	45.4	73.2	82.1	2.0	9.3
		48.0	73.5	82.7	2.0	<b>13.4</b>	<b>47.1</b>	<b>74.2</b>	<b>82.9</b>	2.0	<b>9.1</b>
✓	✓	47.4	73.5	82.7	2.0	13.5	46.8	73.2	82.2	2.0	9.2
		<b>49.1</b>	<b>73.9</b>	<b>82.8</b>	2.0	13.5	46.9	73.8	82.1	2.0	9.3

Table 10: Ablation study of  $\ell_{align}$  and  $\ell_{alignsim}$  on MSR-VTT based on S3MA (ViT-B/32).

$\alpha$	$\beta$	Text-to-Video Retrieval					Video-to-Text Retrieval				
		R@1↑	R@5↑	R@10↑	MnR↓	MeanR↓	R@1↑	R@5↑	R@10↑	MnR↓	MeanR↓
0.02	0.01	<b>49.1</b>	73.9	82.8	2.0	13.5	<b>46.9</b>	73.8	82.1	2.0	9.3
0.02	0.02	48.5	73.8	<b>83.2</b>	2.0	14.0	46.3	73.1	82.1	2.0	9.4
0.02	0.05	47.6	72.7	82.4	2.0	14.0	45.8	<b>74.0</b>	82.2	2.0	9.2
0.02	0.1	47.7	72.3	82.9	2.0	13.4	45.3	73.6	<b>83.3</b>	2.0	<b>9.0</b>
0.01	0.01	47.6	74.0	82.7	2.0	13.8	46.7	73.5	82.2	2.0	9.5
0.05	0.01	48.1	73.6	83.1	2.0	<b>13.2</b>	46.3	72.9	82.7	2.0	9.1
0.1	0.01	47.9	<b>74.2</b>	82.3	2.0	13.3	46.3	73.4	82.5	2.0	9.1

Table 11: Ablation study of  $\alpha$  and  $\beta$  on MSR-VTT based on S3MA (ViT-B/32).

Dense Space				Sparse Space				Text-to-Video Retrieval					Video-to-Text Retrieval				
S-V	S-F	W-V	W-F	S-V	S-F	W-V	W-F	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
✓	✓			✓	✓			<b>49.1</b>	<b>73.9</b>	<b>82.8</b>	2.0	13.5	<b>46.9</b>	73.8	<b>82.1</b>	2.0	<b>9.3</b>
✓	✓	✓	✓	✓	✓	✓	✓	48.3	73.8	82.7	2.0	<b>13.0</b>	46.6	<b>74.1</b>	<b>82.1</b>	2.0	9.4
X-CLIP								46.1	74.3	83.1	2.0	13.2	46.8	73.3	84.0	2.0	9.1

Table 12: Retrieval performance with different similarities on MSR-VTT using S3MA with the base model of ViT-B/32. “S-V”, “S-F”, “W-V”, and “W-F” represent Sentence-Video (coarse-grained), Sentence-Frame (fine-grained), Word-Video (fine-grained), and Word-Frame (fine-grained) similarities.

where  $sim^t$  is the similarity between sentence representations and concepts. The results of this comparison can be found in Table 9. The experimental results show that with the “anchor”, S3MA can better align different modalities as R@1, R@5, and R@10 on text-to-video retrieval and R@1 on video-

to-text retrieval have greatly improved, indicating that the supervised (anchor-based) alignment is crucial for better performance of the model.

**Effect of losses and hyperparameter sensitivity.** To further demonstrate the effectiveness of the

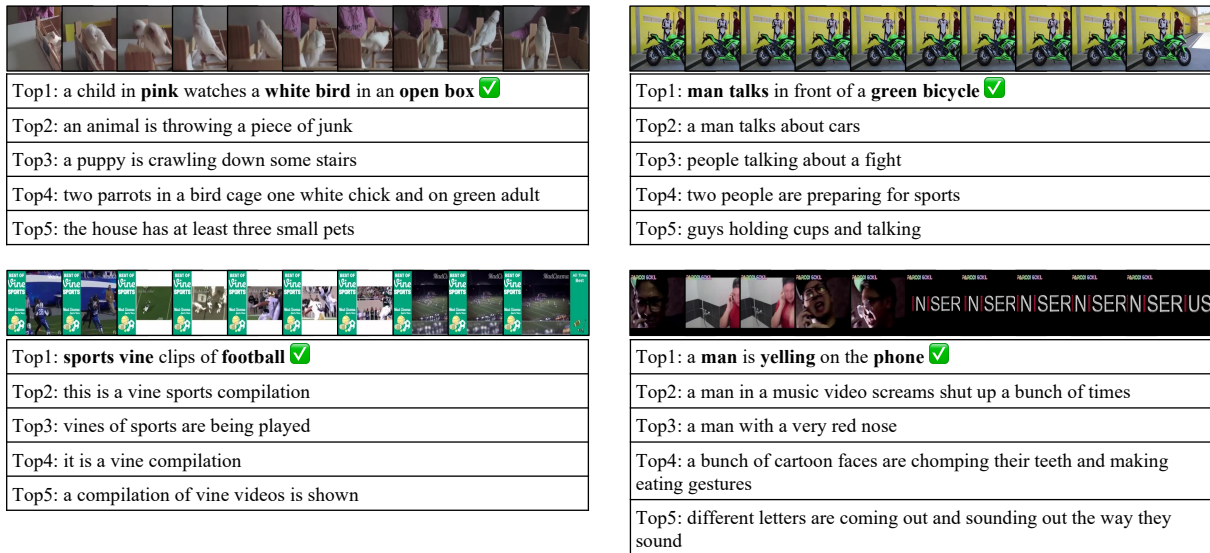


Figure 6: Top-5 video-to-text retrieval results on MSR-VTT.

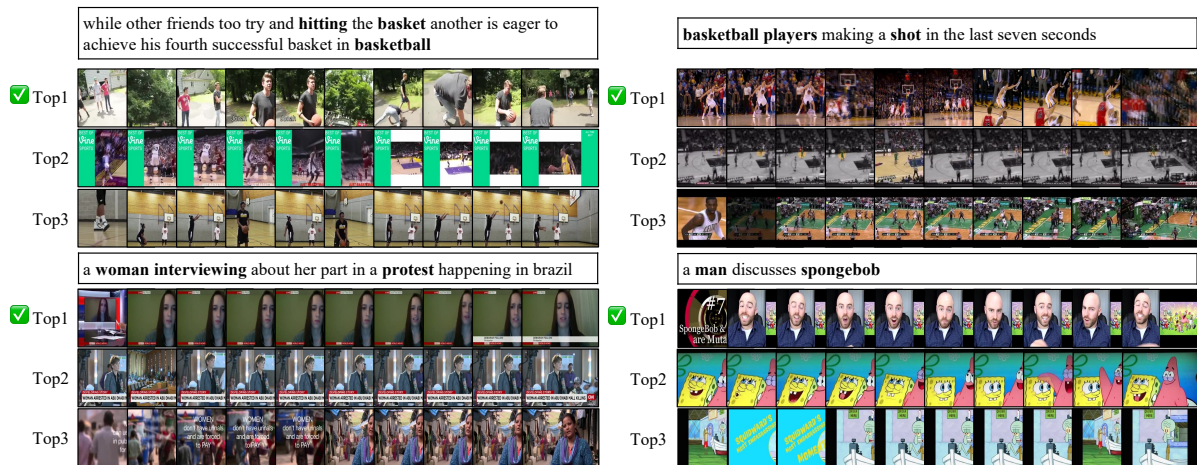


Figure 7: Top-3 text-to-video retrieval results on MSR-VTT.

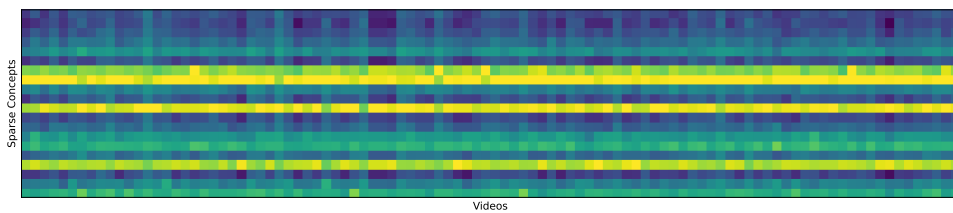


Figure 8: The activation of 20 sparse concepts by 100 randomly selected videos.

two proposed losses designed for aligning different modalities in the shared sparse space, we conduct experiments to compare the performance of these losses. The quantitative results of these experiments are shown in Table 10. The results indicate that adding both losses simultaneously achieves the best performance on the MSR-VTT dataset. When using only one loss, the performance on text-to-

video retrieval is comparable to the method without using both losses on text-to-video retrieval, but outperforms the method without the two losses on video-to-text retrieval. Specifically, when using two losses, R@1 on text-to-video retrieval and video-to-text retrieval is improved by 1.1 and 1.5, respectively. Additionally, all the other metrics, such as R@5 and R@10, are also improved, demon-



strating the power of the two proposed losses in aligning different modalities in the shared sparse space. To gain a better understanding of the sensitivity of S3MA with respect to the two hyperparameters,  $\alpha$  and  $\beta$ , we conduct a series of experiments with different settings of  $\alpha$  and  $\beta$  as shown in Table 10. The results of these experiments demonstrate that, even with varying settings of  $\alpha$  and  $\beta$ , the video-text retrieval performance remains consistent, indicating that the model is robust and not highly sensitive to these hyperparameters. This suggests that S3MA is able to achieve good performance across a wide range of settings for these hyperparameters, making it easy to adjust and optimize for specific use cases. Additionally, this also suggests that S3MA is not overly dependent on precise values of these hyperparameters, and is instead able to leverage the more important underlying features and patterns in the data.

**Are word-level features necessary?** To investigate the necessity of word-level features, we introduce word-level dense and sparse representations, along with word-frame and word-video similarities, into the dense and sparse spaces. The results are presented in Table 12. Notably, we observe a decrease in performance when incorporating word-level contrast in both dense and sparse spaces, indicating possible feature redundancy. Moreover, our approach, which incorporates word-level contrast, can be viewed as an extension of X-CLIP (Ma et al., 2022b) with the shared sparse space. We notice that contrasting representations in the aligned sparse space enhances the retrieval performance of X-CLIP.

#### A.4 Aligning Examples

To show the effectiveness of S3MA, we illustrate some examples of video-to-text and text-to-video retrieval examples in Figures 4, 6 and 7. We notice that S3MA is able to align some important concepts between video and text for precise retrieval. For example, in the bottom-left video-to-text result (Figure 6), the biggest difference between the top 5 retrieved texts is “football”. By precisely capturing “football” in the video, S3MA is able to give higher logits to the sentences that contain “football”. Additionally, in the last (bottom-right) text to video result (Figure 7), we notice that, by understanding “man” and “discuss”, S3MA is able to distinguish the top 3 retrieved videos and select the one in which a man appears. This empirically shows that

S3MA performs well in visual and textual content understanding, benefiting from multi-space and multi-grained similarity.

Moreover, we visualize the activation of sparse concepts by videos in Figure 8. We notice that, some hub sparse concepts are frequently retrieved while some are not retrieved a lot, which might be due to the KNN clustering. Moreover, we notice that the difference between activations from videos are separable.