X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval

Yiwei Ma *
Media Analytics and Computing Lab,
Xiamen University
Xiamen, China
yiweima@stu.xmu.edu.cn

Guohai Xu
DAMO Academy, Alibaba Group
Hangzhou, China
guohai.xgh@alibaba-inc.com

Xiaoshuai Sun †
Media Analytics and Computing Lab,
Xiamen University
Institute of Artificial Intelligence,
Xiamen University
Xiamen, China
xssun@xmu.edu.cn

Ming Yan
DAMO Academy, Alibaba Group
Hangzhou, China
ym119608@alibaba-inc.com

Ji Zhang
DAMO Academy, Alibaba Group
Hangzhou, China
zj12214@alibaba-inc.com

Rongrong Ji
Media Analytics and Computing Lab,
Xiamen University
Institute of Artificial Intelligence,
Xiamen University
Xiamen, China
rrji@xmu.edu.cn

ABSTRACT
Video-text retrieval has been a crucial and fundamental task in multi-modal research. The development of video-text retrieval has been considerably promoted by large-scale multi-modal contrastive pre-training, which primarily focuses on coarse-grained or fine-grained contrast. However, cross-grained contrast, which is the contrast between coarse-grained representations and fine-grained representations, has rarely been explored in prior research. Compared with fine-grained or coarse-grained contrasts, cross-grained contrast calculate the correlation between coarse-grained features and each fine-grained feature, and is able to filter out the unnecessary fine-grained features guided by the coarse-grained feature during similarity calculation, thus improving the accuracy of retrieval. To this end, this paper presents a novel multi-grained contrastive model, namely X-CLIP, for video-text retrieval. However, another challenge lies in the similarity aggregation problem, which aims to aggregate fine-grained and cross-grained similarity matrices to instance-level similarity. To address this challenge, we propose the Attention Over Similarity Matrix (AOSM) module to make the model focus on the contrast between essential frames and words, thus lowering the impact of unnecessary frames and words on retrieval results. With multi-grained contrast and the proposed AOSM module, X-CLIP achieves outstanding performance on five widely-used video-text retrieval datasets, including MSR-VTT (49.3 R@1), MSVD (50.4 R@1), LSMDC (26.1 R@1), DiDeMo (47.8 R@1) and ActivityNet (46.2 R@1).

CCS CONCEPTS
• Information systems → Video search; Novelty in information retrieval.

KEYWORDS
video-text retrieval, multi-grained contrast, attention

ACM Reference Format:
Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, Rongrong Ji. 2022. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547910

1 INTRODUCTION
Video-text retrieval (VTR) is a multi-modal task, which aims to find the most relevant video/text based on the text/video query. With the explosive growth of videos on the Internet, VTR has attracted increasing interests and served as an important role in people’s daily life. Recent years have witnessed the rapid development of VTR, which is supported by a series of pre-training multi-modal models [4, 30, 44], innovative retrieval methods [3, 5, 13–15, 24, 30, 34, 35, 38, 41, 54, 58, 61, 63, 66] and video-text benchmarks [2, 6, 7, 45, 56].

Recently, with great success in large-scale contrastive language-image pre-training, VTR has also achieved great progress. Specifically, with 400M image-text pairs for training, CLIP [44] can embed the images and sentences into the shared semantic space for similarity calculation. Furthermore, CLIP4Clip [38] transfers the image-text knowledge of CLIP to the VTR task, resulting in significant performance improvements on several video-text retrieval datasets. However, CLIP and CLIP4Clip embed the whole sentence and image/video into textual and visual representations, thus lacking the

*Work is done during internship at Alibaba Group. †Corresponding Author.
Figure 1: X-CLIP aims for improving video-text retrieval performance via multi-grained contrastive learning, including fine-grained (frame-word), coarse-grained (video-sentence) and cross-grained (video-word, sentence-frame) contrast. The transparency of words and frames represents the degree of relevance to query.

ability to capture fine-grained interactions. To this end, some previous works [29, 59] propose fine-grained contrastive frameworks, which consider the contrast between each word of the sentence and each frame of the video. Moreover, TACo [57] introduces token-level and sentence-level loss to consider both fine-grained and coarse-grained contrast. Although they have shown promising advances on the VTR task, cross-modality semantic contrast still needs to be systematically explored.

As shown in Fig. 1, a video is composed of multiple frames, and a sentence consists of several words. Video and sentence are usually redundant, which may contain some unnecessary frames or unimportant words. Concretely, given a specific video or sentence query, unnecessary frames or unimportant words refer to the candidates with low relevance to the query (i.e., light-colored frames and words in Fig. 1). However, most current works mainly focus on coarse-grained contrast [38, 44], fine-grained contrast [29, 59] or both [57], which are inefficient in filtering out these unnecessary frames and words. Specifically, coarse-grained contrast calculates the similarity between video-level and sentence-level features, and fine-grained contrast calculates the similarity between frame-level and word-level features. To this end, we ask: How to effectively filter out unnecessary information during retrieval? To answer this question, we propose the cross-grained contrast, which calculates the similarity score between the coarse-grained features and each fine-grained feature. As shown in Fig. 1, with the help of the coarse-grained feature, unimportant fine-grained features will be filtered out and important fine-grained features will be up-weighted. However, challenges in cross-grained contrast arise from aggregating similarity matrices to instance-level similarity scores. A naive and easy method is to use Mean-Max strategy [25, 26, 47, 59] to calculate the instance-level similarity score after obtaining the similarity matrix. However, the conventional Mean-Max strategy is not conducive to filtering out the unnecessary information in videos and sentences during retrieval.

Based on the above analysis, in this paper, we propose an end-to-end multi-grained contrast model, namely X-CLIP, for video-text retrieval. Specifically, X-CLIP first adopts modality-specific encoders to generate multi-grained visual and textual representations and then considers multi-grained contrast of features (i.e., video-sentence, video-word, sentence-frame, and frame-word) to obtain multi-grained similarity scores, vectors, and matrices. To effectively filter out the unnecessary information and obtain meaningful instance-level similarity scores, the AOSM module of X-CLIP conducts the attention mechanism over the similarity vectors/matrices. Different from the conventional Mean-Max strategy, our proposed AOSM module dynamically considers the importance of each frame in the video and each word in the sentence, so the adverse effects of unimportant words and unnecessary frames on retrieval performance are reduced.

To validate the effectiveness of our proposed X-CLIP, we conduct extensive experiments on five widely-used video-text retrieval benchmarks and achieve significantly better performance than previous approaches. Specifically, our X-CLIP achieves 49.3 R@1 on MSR-VTT (i.e., 6.3% relative improvement, 2.9% absolute improvement over the previous state-of-the-art approach). Besides, our proposed X-CLIP achieves 50.4 R@1, 26.1 R@1, 47.8 R@1, 46.2 R@1 on the MSVD, LSMDC, DiDeMo and ActivityNet datasets, respectively, which outperforms the previous SOTA method by +6.6% (+3.1%), +11.1% (+2.6%), +6.7% (+3.0%), +3.8% (+1.7%) on relative (absolute) improvement.

2 RELATED WORKS

2.1 Vision-Language Pre-Training

With the success of self-supervised pre-training such as BERT [12] in NLP, vision-language pre-training on large-scale unlabeled cross-modal data has attracted growing attention [23, 31–33, 37, 44, 50, 51, 55, 60]. One line of work such as LXMERT [51], OSCAR [33] and ALBEF [31] focuses on pre-training on enormous image-text pairs data, and obtains significant improvement in a variety of vision-and-language tasks. To better cope with the image-text retrieval tasks, contrastive language-image pre-training methods such as CLIP [44], ALIGN [23] and WenLan [19] have been proposed, by leveraging billion-scale image-text pairs data from the web with a dual-stream Transformer. Due to the great advantage of CLIP for visual representation learning, some recent work such as CLIP4Clip [38] has also begun to transfer the knowledge of CLIP to video-text retrieval tasks and obtained new state-of-the-art results. The other line of work such as VideoBERT [50], HERO [32] and Frozen in Time [4] directly collects video-text pairs data for video-language pre-training, by further considering the temporal information in videos. However, the scale of the video-language pre-training dataset is much smaller than image-text pre-training since the process of video-text dataset collection is much more expensive. In this work, we follow the line of CLIP4Clip [38], which enhances video-text retrieval by borrowing the ability of visual representation learning from contrastive image-text pre-training. Different from CLIP4Clip [38], we design a multi-grained video-text alignment function to better align the video-text semantics.

2.2 Video-Text Retrieval

Video-text retrieval is a popular but challenging task, which involves cross-modal fusion of multiple modalities and additional

---

MM ’22, October 10–14, 2022, Lisboa, Portugal

Yiwei Ma et al.

---

MM ’22, October 10–14, 2022, Lisboa, Portugal

Yiwei Ma et al.

understanding of temporal information in videos. Traditional video-text retrieval methods tend to design task-specific or modality-specific fusion strategies for cross-modal learning from offline extracted video and text features [15, 16, 20, 28, 34, 42, 62], including face recognition/object recognition/audio processing. However, they are limited by the pre-extracted single modal features, since these features are not properly learnt for the target downstream tasks. Recently, the paradigm of end-to-end video-text retrieval by training models directly from raw video/text has gained large popularity. For example, MIL-NCE [40] adopts Multiple Instance Learning and Noise Contrastive Estimation for end-to-end video representation learning, which addresses visually misaligned narrations from uncurated videos. ClipBERT [30] proposes to sparsely sample video clips for end-to-end training to obtain clip-level predictions, while Frozen in Time [4] uniformly samples video frames and conducts end-to-end training on both image-text and video-text pairs data. CLIP4Clip [38] transfers the knowledge of CLIP to end-to-end video-text retrieval and investigates three similarity calculation approaches for video-sentence contrastive learning. However, cross-grained (i.e., video-word and sentence-frame) contrast is also critical, which has rarely been explored in previous works. We propose the first work of multi-grained contrastive learning for end-to-end video-text retrieval, by considering all the video-sentence, video-word, sentence-frame, and frame-word contrasts.

### 2.3 Multi-Grained Contrastive Learning

Recently, contrastive learning [8–10, 18] has been a popular topic in deep learning community. CLIP [44] implements the idea of contrastive learning based on a large number of image-text pairs, achieving outstanding performance on several multi-modal downstream tasks [17, 21, 22, 39, 64, 65]. To achieve fine-grained contrastive learning, FILIP [59] contrasts the patch in the image with the word in the sentence, achieving fine-grained semantic alignment. TACo [57] proposes token-level and sentence-level losses to include both fine-grained and coarse-grained contrasts. Although contrastive learning has been widely used in multi-modal pre-training, cross-grained contrast has rarely been explored in previous works, which is also critical for semantic alignment. Therefore, we propose a multi-grained contrastive learning method for video-text retrieval, which aims to achieve multi-grained semantic alignment.

### 3 METHODOLOGY

In this section, we elaborate each component of our proposed X-CLIP, whose architecture is shown in Fig. 2. Specifically, we first introduce how to extract the multi-grained visual and textual representations in Sec. 3.1. We then explain the multi-grained contrastive learning based on these feature representations in Sec. 3.2, which aims to obtain multi-grained contrast scores, vectors, and matrices. We also introduce how to aggregate the similarity vectors/matrices to the instance-level similarity score in Sec. 3.3. Finally, we describe the similarity calculation and objective function for video-text retrieval in Sec. 3.4 and 3.5, respectively.

### 3.1 Feature Representation

#### 3.1.1 Frame-level Representation

For a video \( v \in \bar{V} \), we first sample video frames using the sampling rate of 1 frame per second (FPS). Frame encoder is used to process these frames to obtain frame-level features, which is a standard vision transformer (ViT) with 12 layers. Following the previous work [38], we initialize our frame encoder with the public CLIP [44] checkpoints. The architecture of ViT is the same as the transformer [52] encoder in natural language processing (NLP), except ViT introduces a visual tokenization process to convert video frames into discrete token sequences. The discrete token sequence, which is prepended with a [CLS] token, is then fed into the Transformer of ViT. The [CLS] tokens from the last layer are extracted as the frame-level features \( \hat{o}_{(i,j)} \in \bar{V} \).

#### 3.1.2 Visual Representation

Given the video-level representation \( \hat{o}_{(i,j)} \in \bar{V} \) are extracted from separate frames, without considering the interaction among frames. Therefore, we further propose a temporal encoder with temporal position embedding \( \mathbf{P} \), which is a set of predefined parameters, to model the temporal relationship. To be specific, the temporal encoder is also a standard transformer with 3 layers, which can be formulated as:

\[
\mathbf{V}_i = \text{TransEnc} \left( \hat{V}_i + \mathbf{P} \right),
\]

where \( \mathbf{V}_i = \{ v_{(i,1)}, v_{(i,2)}, \ldots, v_{(i,n)} \} \) is the final frame-level (fine-grained) visual features for the video \( \hat{o}_i \). To obtain video-level (coarse-grained) visual feature \( v_i' \in \mathbb{R}^{\text{dim}} \), all frame-level features of the video \( v_i \) are averaged, which can be formulate as:

\[
v_i' = \frac{1}{n} \sum_{j} v_{(i,j)}.
\]

#### 3.1.3 Textual Representation

Given a sentence, we directly use the text encoder of CLIP to generate the textual representation, which is also initialized by the public checkpoints of CLIP [44]. Specifically, it is a transformer encoder, which consists of multi-head self-attention and feed-forward networks. The transformer consists of 12 layers and 8 attention heads. The dimension of the query, key, and value features is 512. The tokenizer used in the experiment is lower-cased byte pair encoding (BPE) [48] with a 49,152 vocab size. Before being fed into the text encoder, the textual token sequence is padded with [BOS] and [EOS] at the beginning and end, respectively. The sentence-level (coarse-grained) textual feature \( t_i' \in \mathbb{R}^{\text{dim}} \) and word-level (fine-grained) textual features \( T_i = \{ t_{(i,1)}, t_{(i,2)}, \ldots, t_{(i,m)} \} \) are the outputs of the [EOS] token, and corresponding word tokens from the final layer of text encoder, where \( m \) is the length of the sentence.

### 3.2 Multi-Grained Contrastive Learning

Previous VTR works [29, 38] focus on fine-grained and coarse-grained contrastive learning, which include video-sentence and frame-word contrasts. However, as explained in Sec. 1, cross-grained (i.e., video-word and sentence-frame) contrast is explicit to filter out the unnecessary information in the video and sentence. Therefore, different from previous works [29, 38, 59], which only focus single-grained contrast, X-CLIP is a multi-grained contrastive framework for VTR.

#### 3.2.1 Video-Sentence Contrast

Given the video-level representation \( v' \in \mathbb{R}^{\text{dim}} \) and sentence-level representation \( t' \in \mathbb{R}^{\text{dim}} \), we use matrix multiplication to evaluate the similarity between video...
and sentence, which can be formulated as:

$$S_{V \rightarrow S} = (\sigma')^T (t'),$$  \hspace{1cm} (3)$$

where $S_{V \rightarrow S} \in \mathbb{R}^1$ is the video-sentence similarity score.  

### 3.2.2 Video-Word Contrast

For the given video-level representation $\sigma' \in \mathbb{R}^{dim}$ and word-level representation vector $t \in \mathbb{R}^{maxdim}$, we use matrix multiplication to calculate the similarity between the video representation and each word representation, which can be represented as follows:

$$S_{V \rightarrow W} = (Te')^T,$$  \hspace{1cm} (4)$$

where $S_{V \rightarrow W} \in \mathbb{R}^{1 \times m}$ is the similarity vector between video and each word in the sentence, $m$ is the length of the sentence.

### 3.2.3 Sentence-Frame Contrast

Similar to Video-Word Contrast, we can calculate the similarity between the sentence representation $t' \in \mathbb{R}^{dim}$ and each frame representation $V \in \mathbb{R}^{n \times dim}$ based on matrix multiplication, which can be formulated as follows:

$$S_{F \rightarrow S} = Vt',$$  \hspace{1cm} (5)$$

where $S_{F \rightarrow S} \in \mathbb{R}^{n \times 1}$ is the similarity vector between the sentence and each frame of a video, $n$ is the number of frames in the video.

### 3.2.4 Frame-Word Contrast

The fine-grained similarity matrix between word representations and frame representations can be also obtained using the matrix multiplication:

$$S_{F \rightarrow W} = VT^T,$$  \hspace{1cm} (6)$$

where $S_{F \rightarrow W} \in \mathbb{R}^{n \times m}$ is the fine-grained similarity matrix, $n$ and $m$ are the number of frames and words, respectively.

### 3.3 Attention Over Similarity Matrix (AOSM)

To obtain the instance-level similarity, we fuse the similarity vector/matrix in Eq. 4, Eq. 5 and Eq. 6. As discussed in Sec. 1, Mean-Max strategies [25, 26, 47, 59] ignore the importance of different frames and words. To address this issue, we propose the Attention Over Similarity Matrix (AOSM) module, where scores in similarity vectors/matrices will be given different weights during aggregation.

Specifically, given the similarity vectors $S_{V \rightarrow W} \in \mathbb{R}^{1 \times m}$ and $S_{F \rightarrow S} \in \mathbb{R}^{n \times 1}$, we first use Softmax to obtain the weights for the similarity vector, where scores for the fine-grained features related to the query will be given high weights. Then, we aggregate these similarity scores based on the obtained weights, which can be formulated as follows:

$$S'_{V \rightarrow W} = \sum_{i=1}^{m} \frac{\exp(S_{V \rightarrow W(i,i)/\tau})}{\sum_{j=1}^{m} \exp(S_{V \rightarrow W(j,j)/\tau})} S_{V \rightarrow W(i,i)},$$  \hspace{1cm} (7)$$

$$S'_{F \rightarrow S} = \sum_{i=1}^{n} \frac{\exp(S_{F \rightarrow S(i,i)/\tau})}{\sum_{j=1}^{n} \exp(S_{F \rightarrow S(j,j)/\tau})} S_{F \rightarrow S(i,i)},$$  \hspace{1cm} (8)$$

where $\tau$ is the temperature parameter of Softmax.

Since the fine-grained similarity matrix $S_{F \rightarrow W} \in \mathbb{R}^{n \times m}$ contains the similarity scores of $n$ frames and $m$ words, we perform attention operations on the matrix twice. The first attention aims to get fine-grained video-level and sentence-level similarity vectors, which can be formulated as follows:

$$S_{vid} = \sum_{i=1}^{n} \frac{\exp(S_{F \rightarrow W(i,i)/\tau})}{\sum_{j=1}^{n} \exp(S_{F \rightarrow W(j,j)/\tau})} S_{F \rightarrow W(i,i)},$$  \hspace{1cm} (9)$$
We use 1,200, 100, 670 videos for training, validating, and testing. We use the average value as the fine-grained similarity score:

\[ S_{\text{sen}} = \frac{1}{m} \sum_{i=1}^{m} \frac{\exp(S_{F-W}(s_i/t))}{\sum_{j=1}^{n} \exp(S_{F-W}(s_j/t))} S_{F-W}(s_i, t), \]  

where \( * \) represents all content in the dimension, \( S_{\text{vid}} \in \mathbb{R}^{1 \times m} \) and \( S_{\text{sen}} \in \mathbb{R}^{n \times 1} \) are the video-level and sentence-level similarity vector, respectively. Specifically, \( S_{\text{vid}} \in \mathbb{R}^{1 \times m} \) shows the similarity score between the video and \( m \) words in the sentence. \( S_{\text{sen}} \in \mathbb{R}^{n \times 1} \) represents the similarity score between the sentence and \( m \) frames in the video.

To obtain fine-grained instance-level similarity scores, we conduct the second attention operation on the video-level vector \( S_{\text{vid}} \in \mathbb{R}^{1 \times m} \) and sentence-level similarity vector \( S_{\text{sen}} \in \mathbb{R}^{n \times 1} \), which can be represented as follows:

\[ S'_{\text{vid}} = \frac{1}{m} \sum_{i=1}^{m} \frac{\exp(S_{\text{vid}}(1, i))}{\sum_{j=1}^{n} \exp(S_{\text{vid}}(1, j))} S_{\text{vid}}(1, i), \]

\[ S'_{\text{sen}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\exp(S_{\text{sen}}(i, 1))}{\sum_{j=1}^{n} \exp(S_{\text{sen}}(i, j))} S_{\text{sen}}(i, 1), \]

where \( S'_{\text{vid}} \in \mathbb{R}^{1} \) and \( S'_{\text{sen}} \in \mathbb{R}^{1} \) are the instance-level similarities.

We use the average value as the fine-grained similarity score:

\[ S_{F-W}' = (S'_{\text{vid}} + S'_{\text{sen}})/2. \]

### 3.4 Similarity Calculation

The similarity score \( s(v, t) \) measures the semantic similarity between the two instances. Different from the previous work [38] that only consider the coarse-grained contrast, our proposed X-CLIP adopt multi-grained contrast during retrieval. Therefore, the final similarity score \( s(v, t) \) of X-CLIP contains multi-grained contrastive similarity scores, which can be represented as follows:

\[ s(v, t) = (s_{F-S} + s_{F-W}' + S_{F-W} + S_{F-W}')/4. \]

### 3.5 Objective Function

During training, given a batch of \( B \) video-text pairs, the model will generate a \( B \times B \) similarity matrix. We adopt the symmetric InfoNCE loss over the similarity matrix to optimize the retrieval model, which can be formulated as:

\[ L_{\text{clt}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s(v_i, t_i))}{\sum_{j=1}^{B} \exp(s(v_i, t_j))}, \]

\[ L_{\text{t2v}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s(v_i, t_i))}{\sum_{j=1}^{B} \exp(s(v_j, t_i))}, \]

\[ L = L_{\text{clt}} + L_{\text{t2v}}. \]

### 4 EXPERIMENTS

#### 4.1 Datasets

**MSR-VTT** [56] is a popular video-text retrieval dataset, which contains 10,000 videos and 200,000 captions. The length of videos in this dataset ranges from 10 to 32 seconds. In this paper, we adopt the widely-used ‘Training-9K’ split, where 9,000 videos and 180,000 captions are used for training and the rest are used for testing.

**MSVD** [7] contains 1,970 videos, the duration of which vary from 1 to 62 seconds. Each video is annotated with 40 English captions. We use 1,200, 100, 670 videos for training, validating, and testing.

**LSMDC** [45] is a dataset that contains 118,081 videos and captions. The duration of each video ranges from 2 to 30 seconds. We adopt 109,673, 7,408, and 1,000 videos for training, validating, and testing.

**DiDeMo** [2] contains 10,000 videos and 40,000 captions. Following previous works [4, 30, 35], all captions of a video are concatenated together during video-paragraph retrieval.

**ActivityNet** [6] contains 20,000 YouTube videos, which are annotated temporally. Following previous works [15, 38, 49], all captions of a video are also concatenated together during video-paragraph retrieval for fair comparison.

#### 4.2 Experimental Settings

##### 4.2.1 Implementation Details

We conduct the experiments on 4 NVIDIA Tesla V100 32GB GPUs using the PyTorch library. Following the previous work [38], the text encoder and frame encoder of X-CLIP are initialized by the public CLIP checkpoints. We use the Adam optimizer [27] to optimize the X-CLIP and decay the learning rate using a cosine schedule strategy [36]. Since the parameters of the text encoder and frame encoder are initialized from the public CLIP checkpoints, we adopt different learning rates for different modules. Specifically, the initial learning rate for text encoder and frame encoder is 1e-7, and the initial learning rate for other modules is 1e-4.

We set the max token length, max frame length, batch size, and the training epoch to 32, 12, 300, and 3 for MSR-VTT, MSVD, and LSMDC datasets. Since videos and captions in DiDeMo and ActivityNet are longer and more complex, we set the max token length, max frame length, and the training epoch to 64, 64, and 20. Due to the limitation of GPU memory, we also reduce the batch size of DiDeMo and ActivityNet to 64. We conduct ablation, quantitative and qualitative experiments on the MSR-VTT dataset, it is more popular and competitive compared with other datasets. The base model of X-CLIP is ViT-B/32 if not specified.

##### 4.2.2 Evaluation Protocols

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 [3, 5, 13–15, 30, 34, 35, 38, 41, 61, 63, 66].

#### 4.3 Performance Comparison

We compare X-CLIP against the previous works on MSR-VTT, MSVD, LSMDC, DiDeMo, and ActivityNet. X-CLIP achieves the SOTA results on all five datasets with significant improvements.

For the MSR-VTT dataset, the performance comparison is shown in Tab. 1. By analyzing the table, we gain the following observations:

- Benefiting from the large-scale image-text pre-training, both CLIP4Clip and our model X-CLIP can obtain significant gains in performance compared with all the baselines. The consistent improvements verify that it is important to adopt end-to-end finetuning to realize the full potential of the image-text pre-trained model on video-text retrieval.
- Compared with the strongest competitor (i.e., CLIP4Clip-seqTransf), X-CLIP obtains 49.3 R@1 (6.3% relative improvement, 2.9% absolute improvement) in the text-to-video retrieval task and 48.9 R@1 (7.7% relative improvement, 3.5% absolute improvement) in the video-to-text retrieval task by employing CLIP(ViT-B/16) as a pre-trained model. This can be attributed to that our proposed
Table 1: Retrieval performance comparison to SOTAs on the MSR-VTT dataset.

| Model                                                                 | Text-to-Video Retrieval | Video-to-Text Retrieval |
|----------------------------------------------------------------------|-------------------------|-------------------------|
|                                                                      | R@1↑                    | R@5↑                    | R@10↑              | MdR↑ | MnR↑ | R@1↑             | R@5↑        | R@10↑              | MdR↑      | MnR↑ |
| CE [35]                                                              | 70.9 84.0               | 2.0 16.2                | 43.1 70.5          | 81.2 2.0 | 12.4 |
| MMT [15]                                                             | 42.5 70.8 80.7          | 2.0 16.7                | 42.7 71.0          | 80.4 2.0 | 12.3 |
| AVLnet [46]                                                          | 40.2 71.5 80.5          | 2.0 13.4                | 40.6 69.5          | 79.5 2.0 | 13.6 |
| SSB [42]                                                             | 35.3 73.3 83.0          | 2.0 13.0                | 44.8 73.2          | 82.2 2.0 | 9.6  |
| MDMMT [14]                                                           | 44.3 72.0 82.2          | 2.0 13.7                | 44.3 73.4          | 82.4 2.0 | 10.3 |
| Frozen [4]                                                           | 44.6 72.1 82.0          | 2.0 14.7                | 45.4 73.4          | 82.4 2.0 | 10.7 |
| HIT [34]                                                             | 42.9 71.7 81.5          | 2.0 13.3                | 41.9 71.0          | 80.7 2.0 | 10.1 |
| TT-CE+ [11]                                                          | 46.1 73.0 83.1          | 2.0 13.2                | 46.8 73.3          | 84.0 2.0 | 9.1  |
| X-CLIP (ViT-B/32)                                                    | 49.3 75.8 84.8          | 2.0 12.2                | 48.9 76.8          | 84.5 2.0 | 8.1  |
| X-CLIP (ViT-B/16)                                                    |                        |                        |                      |      |      |                  |            |                   |           |

Table 2: Retrieval performance comparison on MSVD.

| Model                                                                 | Text-to-Video | Video-to-text |
|----------------------------------------------------------------------|---------------|---------------|
|                                                                      | R@1↑ R@5↑ MnR↑ | R@1↑ R@5↑ MnR↑ |
| MultiCues [41]                                                       | 79.7 73.2 | -             | 75.5 10.3 | -             |
| CE [35]                                                              | 87.5 85.8 | -             | 87.3 85.4 | -             |
| SSB [42]                                                             | 54.2 54.2 | 56.9 56.9 | 27.1 55.3 | -             |
| NoiseE [1]                                                           | 40.3 39.3 | 40.3 39.3 | -             | -             |
| CLIP-straight [43]                                                   | 61.4 61.4 | 60.9 60.9 | 59.9 60.9 | -             |
| Frozen [4]                                                           | 61.4 61.4 | 61.4 61.4 | -             | -             |
| TT-CE+ [11]                                                          | 57.0 57.0 | 57.0 57.0 | -             | -             |
| X-CLIP (ViT-B/32)                                                    | 64.7 64.7 | 64.7 64.7 | -             | -             |
| X-CLIP (ViT-B/16)                                                    | 64.7 64.7 | 64.7 64.7 | -             | -             |

Table 3: Retrieval performance comparison on LSMDC.

| Model                                                                 | Text-to-Video | Video-to-Text |
|----------------------------------------------------------------------|---------------|---------------|
|                                                                      | R@1↑ R@5↑ MnR↑ | R@1↑ R@5↑ MnR↑ |
| CLIP-straight [43]                                                   | 56.9 56.9 | -             | 56.9 56.9 | -             |
| Frozen [4]                                                           | 61.4 61.4 | 61.4 61.4 | -             | -             |
| TT-CE+ [11]                                                          | 57.0 57.0 | 57.0 57.0 | -             | -             |
| X-CLIP (ViT-B/32)                                                    | 64.7 64.7 | 64.7 64.7 | -             | -             |
| X-CLIP (ViT-B/16)                                                    | 64.7 64.7 | 64.7 64.7 | -             | -             |

cross-grained contrast and the AOSM module are critical to reducing the bad effects of unnecessary frames and unimportant words.

- Compared to all the other state-of-the-arts, our model with ViT-B/16 achieves the best performance in all metrics. Surprisingly, our model with the ViT-B/32 can even achieve comparable performance to CLIP4Clip with ViT-B/16, which again demonstrates the effectiveness and superiority of multi-grained contrast and the AOSM module.

We also further validate the generalization of X-CLIP on MSVD, LSMDC, DiDeMo and ActivityNet in Tab. 2 - 5. It is worth noting that, in all variants of CLIP4Clip, we only report the performance of CLIP4Clip-MeanP and CLIP4Clip-seqTransf, because they perform better than the other two variants in consideration of experience in the previous work [38] and performance comparison in Tab. 1. By analyzing these tables, we can observe that X-CLIP also achieves significant improvement on these datasets for text-to-video and video-to-text retrieval tasks. Specifically, for the text-to-video retrieval task, X-CLIP outperforms the CLIP4Clip with ViT-B/16 on R@1 by +6.6% (+3.1%), +11.1% (+2.6%), +6.7% (+3.0%), +3.8% (+1.7%) relative (absolute) improvement on aforesaid four datasets respectively.

- With the number of contrastive modules increasing, the retrieval performance tends to be higher. When X-CLIP is equipped with all contrastive modules, the best retrieval performance can be achieved. This may be because each contrastive module plays a different role in the retrieval task and different contrast modules can promote each other to achieve better retrieval results.

4.4 Ablation Study

To fully examine the impact of different contrastive modules, we conduct an ablation study to compare different variants of X-CLIP. As shown in Tab. 6, we gain two important observations:

- We conduct an ablation study to compare different variants of X-CLIP. Specifically, for the text-to-video retrieval task, X-CLIP outperforms the CLIP4Clip with ViT-B/16 on R@1 by +6.6% (+3.1%), +11.1% (+2.6%), +6.7% (+3.0%), +3.8% (+1.7%) relative (absolute) improvement on aforesaid four datasets respectively. For the video-to-text retrieval task, X-CLIP obtains +5.7% (+3.6%), +12.9% (+3.0%), +1.3% (+0.6%), +5.2% (+2.3%) relative (absolute) improvement on R@1. This demonstrates that our proposed X-CLIP can achieve consistent performance improvement on several video-text retrieval datasets. More experimental results are in the supplementary materials.

- As shown in Tab. 6, we gain two important observations: X-CLIP can achieve consistent performance improvement on several video-text retrieval datasets. More experimental results are in the supplementary materials.
Table 4: Retrieval performance comparison on DiDeMo.

| Model                  | R@1 | R@5 | M@1 | R@1 | R@5 | M@1 |
|------------------------|-----|-----|-----|-----|-----|-----|
| SVT [53]               | 11.9| 53.6| -   | 13.2| 33.6| -   |
| FSE [63]               | 13.9| 36.0| -   | 13.1| 33.9| -   |
| CE [57]                | 16.1| 41.1| 43.7| 15.6| 40.9| 42.4|
| ClipBERT [30]          | 20.4| 48.0| -   | -   | -   | -   |
| Frozen [4]             | 34.6| 65.0| -   | -   | -   | -   |
| TT-CE+ [11]            | 21.6| 48.6| -   | 21.1| 47.3| -   |
| CLIP4clip-MeanP (ViT-B/32) [38] | 43.4| 70.2| 17.5| 42.5| 70.6| 11.6|
| CLIP4clip-seqTransf (ViT-B/16) [38] | 42.8| 68.5| 18.9| 41.4| 68.2| 12.4|
| CLIP4clip-MeanP (ViT-B/16) [38] | 44.8| 75.1| 13.0| 47.2| 74.0| 10.5|
| CLIP4clip-seqTransf (ViT-B/16) [38] | 44.8| 73.4| 13.5| 44.7| 74.0| 10.6|
| X-CLIP (ViT-B/32)      | 45.2| 74.0| 14.5| 43.1| 72.2| 10.9|
| X-CLIP (ViT-B/16)      | 47.8| 79.3| 12.6| 47.8| 76.8| 10.5|

Table 5: Retrieval performance comparison on ActivityNet.

| Model                  | R@1 | R@5 | M@1 | R@1 | R@5 | M@1 |
|------------------------|-----|-----|-----|-----|-----|-----|
| FSE [63]               | 18.2| 44.8| -   | 16.7| 43.1| -   |
| CE [57]                | 18.2| 47.7| 23.1| 17.7| 46.6| 24.4|
| HSE [53]               | 20.5| 49.1| -   | 18.7| 48.1| -   |
| MMT [15]               | 28.7| 61.4| 16.0| 28.9| 61.1| 17.1|
| SSB [42]               | 29.2| 61.6| -   | 28.7| 60.8| -   |
| HT [14]                | 29.4| 60.7| -   | -   | -   | -   |
| ClipBERT [30]          | 23.1| 49.0| -   | -   | -   | -   |
| TT-CE+ [11]            | 23.5| 57.2| -   | 23.0| 56.1| -   |
| CLIP4clip-MeanP (ViT-B/32) [38] | 40.5| 72.4| 7.4| 42.5| 74.1| 6.6|
| CLIP4clip-seqTransf (ViT-B/32) [38] | 40.5| 72.4| 7.5| 41.6| 73.7| 6.7|
| CLIP4clip-MeanP (ViT-B/16) [38] | 44.0| 73.9| 7.0| 44.1| 74.0| 6.5|
| CLIP4clip-seqTransf (ViT-B/16) [38] | 44.5| 75.2| 6.4| 44.1| 75.2| 6.4|
| X-CLIP (ViT-B/32)      | 44.1| 74.1| 7.7| 43.9| 73.9| 7.5|
| X-CLIP (ViT-B/16)      | 46.2| 75.5| 6.8| 46.4| 75.9| 6.4|

- Our proposed cross-grained contrast can assist fine-grained contrast or coarse-grained contrast to achieve better performance in the retrieval task. Specifically, X-CLIP with the sentence-video contrast module (i.e., Exp1) only achieves 43.0 R@1 in the text-to-video retrieval task. However, when X-CLIP is additionally equipped with cross-grained contrast modules (i.e., Exp8 and Exp9), the performance gets obvious absolute improvements of 2.4% and 1.0% respectively. Similarly, when X-CLIP is only equipped with fine-grained and coarse-grained contrast modules (i.e., Exp10), it achieves 44.8 R@1 in the text-to-video task. However, when it is additionally equipped with cross-grained contrast modules (i.e., Exp13 and Exp14), 1.0% and 0.7% absolute improvement of R@1 can be achieved. Therefore, we conclude that the performance improvement of cross-grained contrast modules in the retrieval task does not conflict with that of coarse-grained and fine-grained contrast modules.

To justify the effectiveness of the proposed AOSM module, we compare our method with the conventional Max-Max and other variants (i.e., Max-Max, Max-Mean and Mean-Mean). As shown in Tab. 7, we observe that the Mean-Mean strategy performs worst. This may be because the Mean-Mean strategy, which applies the same weight to all similarity scores during aggregating, cannot eliminate the adverse effects of unnecessary frames and unimportant words on the retrieval results. The Max-Mean, Mean-Max and Max-Max strategies perform better than the Mean-Mean strategy. This can be attributed to that these strategies adopt the highest similarity during aggregation, so contrast scores between unnecessary frames and unimportant words will be filtered out. However, since these strategies adopt the top-1 similarity score, some important similarity scores will also be ignored. To address this issue, we propose the AOSM module, where all similarity scores will be applied with different weights during aggregation. From Tab.

4.5 Effect of Temperature Parameter

To explore the effect of different $\tau$ in the AOSM module, we also designed a group of experiments by setting different temperature parameters $\tau$ in Softmax. From Tab. 9, we observe that the retrieval performance first improves before reaching the saturation point (i.e., $\tau = 0.01$), and then begins to decline slightly. The main reason may be that when $\tau$ is large, too many noisy similarity scores are considered. On the contrary, if the $\tau$ is small, some important similarity scores may be ignored. Besides, our proposed attention mechanism with different $\tau$ consistently performs better than the Mean-Mean strategy, and the attention mechanism with the optimal $\tau$ outperforms other strategies in all evaluation protocols. This justifies that our proposed attention mechanism helps to strengthen the influence of important similarity scores and weaken the influence of noisy similarity scores, thus achieving better retrieval performance.

4.6 Qualitative Analysis

To qualitatively validate the effectiveness of our proposed X-CLIP, we show some typical video-to-text and text-to-video retrieval examples in Fig. 3 and Fig. 4, respectively. From these retrieval results, we find that X-CLIP could accurately understand the content of sentences and videos. Meanwhile, it is robust for X-CLIP to comprehend complex and similar sentences and videos, which is mainly attributed to the multi-grained contrast of our proposed model. To
Table 6: Retrieval performance with different contrastive granularity on the MSR-VTT dataset.

| ID | Sent-Video | Word-Video | Text-to-Video | Video-to-Text |
|----|------------|------------|---------------|---------------|
| Exp1 | ✓ | | 30.79 | 31.01 |
| Exp2 | ✓ | ✓ | 30.79 | 31.01 |
| Exp3 | ✓ | ✓ | 30.79 | 31.01 |
| Exp4 | ✓ | ✓ | 30.79 | 31.01 |
| Exp5 | ✓ | ✓ | 30.79 | 31.01 |
| Exp6 | ✓ | ✓ | 30.79 | 31.01 |
| Exp7 | ✓ | ✓ | 30.79 | 31.01 |
| Exp8 | ✓ | ✓ | 30.79 | 31.01 |
| Exp9 | ✓ | ✓ | 30.79 | 31.01 |
| Exp10 | ✓ | ✓ | 30.79 | 31.01 |

Table 7: Retrieval performance with different fusion methods for similarity matrices on the MSR-VTT dataset. TE is short for temporal encoder.

| Method | Text-to-Video | Video-to-Text |
|--------|---------------|---------------|
| Max-Max | 44.0 | 72.6 | 13.5 | 44.4 | 72.5 | 9.2 |
| Mean-Mean | 43.2 | 71.2 | 14.8 | 42.5 | 70.2 | 11.4 |
| Mean-Max | 44.4 | 71.1 | 14.9 | 44.2 | 71.7 | 10.2 |
| Max-Mean | 44.9 | 71.3 | 13.5 | 43.8 | 71.8 | 9.4 |
| Attention | 46.1 | 73.0 | 13.2 | 46.8 | 73.3 | 9.1 |

Table 8: Ablation study of temporal encoder on the MSR-VTT dataset. TE is short for temporal encoder.

| Base Model | TE | Text-to-Video | Video-to-Text |
|------------|----|---------------|---------------|
| ViT-B/32 | ✓ | 45.2 | 72.9 | 13.8 | 45.6 | 73.9 | 9.2 |
| ViT-B/16 | ✓ | 45.5 | 72.8 | 13.9 | 45.8 | 73.7 | 9.1 |

Table 9: Retrieval performance with different temperature parameters $\tau$ in Softmax on the MSR-VTT dataset.

| $\tau$ | Text-to-Video | Video-to-Text |
|--------|---------------|---------------|
| 1      | 43.9 | 71.6 | 14.5 | 43.5 | 71.3 | 11.3 |
| 0.1    | 45.2 | 72.2 | 14.0 | 45.3 | 73.1 | 9.3 |
| 0.01   | 46.1 | 73.0 | 13.2 | 46.8 | 73.3 | 9.1 |
| 0.001  | 45.6 | 72.2 | 13.7 | 43.6 | 72.5 | 9.4 |

5 CONCLUSION

In this paper, we present X-CLIP, a novel end-to-end multi-grained contrastive model for video-text retrieval, which first encodes the sentences and videos into coarse-grained and fine-grained representations, and conducts fine-grained, coarse-grained, and cross-grained contrasts over these representations. The multi-grained contrast and the AOSM module of X-CLIP help to reduce the negative effects of unnecessary frames and unimportant words during retrieval. Significant performance gains on five popular video-text retrieval datasets demonstrate the effectiveness and superiority of our proposed model.

ACKNOWLEDGMENTS

This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, and No. 62002305), Guangdong Basic and Applied Basic Research Foundation (No.2019B151520049), and the National Science Foundation of Fujian Province of China (No.2021J01002). This work was supported by Alibaba Group through Alibaba Research Intern Program.  

Figure 4: Top-3 text-to-video retrieval results on MSR-VTT. The number in parentheses is the similarity score.

be specific, as shown in the first example in Fig.3, although the top-3 retrieved sentences are similar, our proposed X-CLIP can still choose the correct sentence by understanding the details of sentences and videos. Similarly, as shown in the first example in Fig.4, all top-3 retrieved videos describe the same cartoon, while “squid” does not appear in the second and third videos. Due to the multi-grained contrast, X-CLIP performs well in visual and textual content understanding, so it can retrieve the correct video.
