Unified Coarse-to-Fine Alignment for Video-Text Retrieval

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, Mohit Bansal
UNC Chapel Hill
{ziyangw, ylsung, fengchan, gedas, mbansal}.cs.unc.edu

Abstract

*The canonical approach to video-text retrieval leverages a coarse-grained or fine-grained alignment between visual and textual information. However, retrieving the correct video according to the text query is often challenging as it requires the ability to reason about both high-level (scene) and low-level (object) visual clues and how they relate to the text query. To this end, we propose a **U**nified **Co**arse-to-**fi**ne **A**lignment model, dubbed* UCoFiA. *Specifically, our model captures the cross-modal similarity information at different granularity levels. To alleviate the effect of irrelevant visual clues, we also apply an Interactive Similarity Aggregation module (ISA) to consider the importance of different visual features while aggregating the cross-modal similarity to obtain a similarity score for each granularity. Finally, we apply the Sinkhorn-Knopp algorithm to normalize the similarities of each level before summing them, alleviating over- and under-representation issues at different levels. By jointly considering the cross-modal similarity of different granularity,* UCoFiA *allows the effective unification of multi-grained alignments. Empirically,* UCoFiA *outperforms previous state-of-the-art CLIP-based methods on multiple video-text retrieval benchmarks, achieving* 2.4%, 1.4% *and* 1.3% *improvements in text-to-video retrieval R@1 on MSR-VTT, Activity-Net, and DiDeMo, respectively. Our code is publicly available at* https://github.com/Ziyang412/UCoFiA.

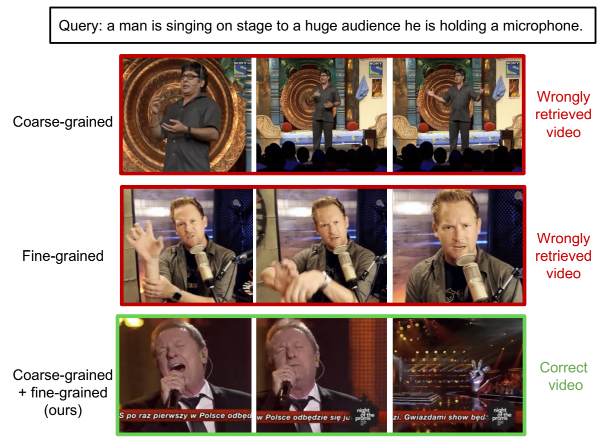

Figure 1. Comparison of the retrieved video from different level cross-modal alignments given a specific query. The coarse-grained alignment (the first row) only observes the high-level scene information and overlooks the detailed information like "microphone". The fine-grained alignment (the second row) does capture the detailed information but ignores the high-level scene information like "stage with the huge audience". To this end, our work (the last row) combines the strengths of coarse-grained and fine-grained alignments by a Unified Coarse-to-fine Alignment model (UCoFiA).

1. Introduction

The fields of computer vision and natural language processing have both seen significant progress in recent years. Thus, the cross-modal alignment [1, 64, 28, 44, 61, 56, 48, 47, 49], which involve developing techniques to connect these two domains, has seen considerable attention and progress. As a direct application of cross-modal alignment, the video-text retrieval task aligns video candidates with text queries to identify the most relevant videos, and the standard practice [50, 62, 63] is to align the video and text features extracted by vision and language encoders. Recently, the emergence of large-scale image-text pretrained models prompted several methods [39, 40, 35] to utilize CLIP [45] image and text encoder to achieve strong performance on many video-text retrieval benchmarks. As a direct extension of CLIP, Luo *et al*. [39] proposed temporal fusion modules to aggregate the features of different video frames and then perform the cross-modal alignment on video and text features. Later, to capture more correspondences between video and text, several works [35, 20, 18, 11] propose to conduct the alignment between frame and text features. Ma *et al*. [40] take a step forward and leverage an alignment between frame and word features for more detailed information.

Although the aforementioned methods have achieved impressive results, they only rely on high-level visual information (frame and video) to perform the cross-modal alignment. This coarse-grained alignment only captures the high-level visual clues (scene, action, etc) that connect to the text query. As shown in the first row of Figure 1, the coarse-grained alignment only captures the scene of the stage with the huge audience" and the action of "singing (possibly talking)", thus leading to the incorrect retrieval result. On the other hand, Zou *et al*. [66] build a fine-grained alignment between patch tokens from the video and word tokens from the text query. As illustrated in the second row of Figure 1, the fine-grained alignment does capture the detailed information like "microphone", but it might overlook high-level clues like scene information ("stage with the huge audience"). These results reveal that video-text retrieval requires an understanding of the both high-level and low-level correspondence between text and video. Thus, in this work, we aim to jointly consider coarse-grained and fine-grained cross-modal alignment and how to unify them to get the correct answer (as shown in the last row of Figure 1).

To this end, we propose UCoFiA, a **U**nified **Co**arse-to-**fi**ne **A**lignment model for video-text retrieval. Our approach aims to capture the multi-grained similarity between the text and video by performing alignment at different granularity. We begin with a coarse-grained alignment between the entire video and the query sentence (video-sentence). Next, we perform frame-sentence alignment by matching individual video frames and the query sentence. Finally, we conduct a fine-grained alignment between the video patches and query words (patch-word).

However, while this multi-grained information provides richer, more diverse detailed information, it also brings significant irrelevant information to the cross-modal alignment. For instance, several frames in the video might not contain information related to the query, and some patches in a frame might only correspond to the background information unrelated to any subjects in the query. The irrelevant information could impede the model from learning precise cross-modal correspondence. To address these issues, we first propose an Interactive Similarity Aggregation module (ISA) that considers the importance of different visual features while aggregating the cross-modal similarity to obtain a similarity score for each granularity. For frame-sentence alignment, our ISA module jointly considers the cross-modal similarity and the interaction of frame features while aggregating the frame-sentence similarity. Compared to the previous methods [35] that ignore the temporal clues between video frames, our ISA module can better capture the important information within continuous video frames. Note that the ISA module is a general similarity aggregation approach regardless of the feature granularity, and we further extend it to a bidirectional ISA module for patch-word alignment.

Next, once we obtain the similarity score for each level of alignment, we can sum them to one score as the final retrieval similarity. However, we find that similarity scores across different videos are highly imbalanced, and we empirically show that correcting this imbalance before summation improves the performance. Concretely, sometimes the sum of retrieval similarities between one specific video and all texts (we called this term marginal similarity) might be much higher than that of the other videos, meaning that this video is over-represented and will lower the probability of the other video being selected. To address this, we utilize the Sinkhorn-Knopp algorithm [14] to normalize the similarity scores and make sure the marginal similarities for different videos are almost identical so that each video has a fair chance to be selected after normalization. We then unify the scores of different levels by performing the algorithm separately on the similarities of different levels and summing them together.

We validate the effectiveness of our UCoFiA model on diverse video-text retrieval benchmarks. Specifically, UCoFiA achieve a text-to-video retrieval R@1 of 49.4% on MSR-VTT [57] and 45.7% on ActivityNet, thus, outperforming the current state-of-the-art CLIP-based methods by 2.4% and 1.4%, respectively.

2. Related Work

Video-text Retrieval. Video-text retrieval [62, 32, 13, 59, 53, 31, 9, 1, 64, 8, 54, 38, 12, 46] is a fundamental topic in the vision-language domain and has attracted significant research attention. To retrieve the correct video candidate given the text query, it is crucial to align the features of the related video and text sample together. To this end, early works in video-text retrieval [50, 62, 63] focus on designing fusion mechanisms for the alignment between pre-extracted and frozen video and text features. Later, ClipBERT [28] proposes a sparse sampling strategy on video data to accomplish end-to-end training and apply image-text pretraining for video-text retrieval. Afterward, Bain *et al*. [3] utilize a curriculum learning schedule to accomplish joint image and video end-to-end training on cross-modal data. With the great success of large-scale image-text pretraining model CLIP [44], several works [39, 4, 18, 6, 23] utilize the powerful CLIP encoder for video-text retrieval tasks and achieve state-of-the-art results with an efficient training paradigm. Thus, in this work, we also use CLIP as our image-text backbone to enable a fair comparison with existing methods.

Moreover, most cross-modal alignment approaches can be divided into two categories: coarse-grained alignment [29, 16, 22, 37, 17, 27, 30, 10, 55, 65] which leverage the frame-level (or video-level) visual features and fine-grained alignment [66, 26, 41, 60, 21] which utilize the information of patch within each video frames. Recently, several coarse-grained CLIP-based methods [39, 4, 20] use frame aggregation strategies to convert frame features to video features and perform coarse-grained alignment between the video and query features. Follow-up works [40, 35] investigate various similarity calculation schemes for better cross-modal alignment. TS2-Net [35] designs a cross-modal alignment model between the frame feature and sentence feature of the text query. X-CLIP [40] utilizes a cross-grained alignment between coarse-grained video features and text features, including video-sentence, video-word, frame-sentence, and frame-word contrast. However, the coarse-grained alignment fails to capture detailed correspondence to due the limited information within high-level features. To this end, TokenFlow [66] proposes a fine-grained alignment function for token-wise similarity calculation. The fine-grained alignment does capture more subtle correspondence between text and video, but it could overlook the high-level information like scene and action. In this work, we aim to combine the advantage of both coarse-grained and fine-grained alignment to better capture the correspondence between text query and video candidates.

Normalization for Video-text Retrieval. To retrieve the most relevant video candidate given a text query, common video-text retrieval models compute a similarity matrix between video and text input and retrieve the candidate with the highest similarity. Several previous works [11, 5, 42] focus on the normalization of this similarity matrix to improve the performance. CAMoE [11] introduces a Dual Softmax Loss (DSL) as a reviser to correct the similarity matrix and achieve the dual optimal match. Later, NCL [42] reveals that cross-modal contrastive learning suffers from incorrect normalization of the sum retrieval probabilities of each text or video instance and proposes Normalized Contrastive Learning that computes the instance-wise biases that properly normalize the sum retrieval probabilities. Empirically, we find that the logits from the multi-level similarity matrix are imbalanced and make some videos and texts over- or under-representative. To mitigate the issue, we propose to balance the multi-level alignments by separately normalizing each similarity matrix and aggregating the normalized matrix for better retrieval prediction.

3. Methodology

In this selection, we present our proposed UCoFiA model. As shown in Figure 2, UCoFiA consists of four components: (1) text and video encoders, (2) coarse-to-fine alignment module, (3) Interactive Similarity Aggregation module, (4) and multi-granularity unification module with the Sinkhorn-Knopp algorithm. First, the video and text encoders extract multi-grained visual and textual features. Afterward, multi-grained features are fed into a coarse-to-fine alignment module that calculates the different levels of cross-modal similarity. Then, the Interactive Similarity Aggregation module fuses the similarity vector (or matrix, depending on the input type) and obtains the similarity score for each granularity. Finally, the multi-granularity unification module aggregates the similarity scores from all granularity and obtains the final unified similarity score for retrieval prediction. Below, we discuss each of these components in more detail.

3.1. Feature Extraction

Text Encoder. Given a text query T (we prepend a [EOS] token to T), we leverage the CLIP [44] text encoder $\mathcal{F}_t$ to output the word feature w, where $w = \mathcal{F}_t(T) \in \mathbb{R}^{L_t \times C}$, L_t denotes the length of word sequences, and C denotes the dimension of the word feature. Then we take the representation of the [EOS] token as the sentence feature $s \in \mathbb{R}^C$.

Video Encoder. Following [39, 35], we utilize the pretrained CLIP visual encoder [15] ($\mathcal{F}_v$) to extract the visual features of each video. A video with N frames can be denoted as $V = [F_1, F_2, ..., F_N]$. Given the n-th frame of the video F_n, we divide it into disjoint patches, prepend a [CLS] token to it, and use the vision encoder $\mathcal{F}_v$ to obtain the patch representation p_n, where $p_n = \mathcal{F}_v(F_n) \in \mathbb{R}^{M \times C}$, M denotes the number of the visual patches within a video frame. Note that both textual and visual tokens are embedded in the same dimension C. Then we take the [CLS] representation from each frame and combine them together to get the frame representation $f = [f_1; f_2; ...; f_N], f \in \mathbb{R}^{N \times C}$. Since we feed frames to ViT separately for better efficiency, the vision encoder cannot learn the temporal information across different frames. To enable the model to learn temporal information with minimal cost, inspired by [35], we adopt a token shift module, which shifts the whole spatial token features back and forth across adjacent frames, in the last two transformer blocks of the ViT model.

3.2. Coarse-to-fine Alignment

Our proposed coarse-to-fine alignment module calculates the cross-modal similarity from multi-grained visual and textual inputs to address the weakness of only considering either coarse alignment or fine-grained alignment (as shown in Figure 1). First, we adopt a video-sentence alignment to obtain the similarity score of video and sentence features. Then, we leverage a frame-sentence alignment to capture the similarity between each frame and text query and obtain a frame-sentence vector. Lastly, we apply the most fine-grained patch-word alignment to model the similarity between each patch and word representation and obtain a patch-word matrix. Below, we describe each of these alignments in more detail.

Video-sentence Alignment. To obtain the video representation, following [39], we leverage a temporal encoder to aggregate the frame features f and obtain the video feature v. We then compute the cosine similarity between $v \in \mathbb{R}^C$

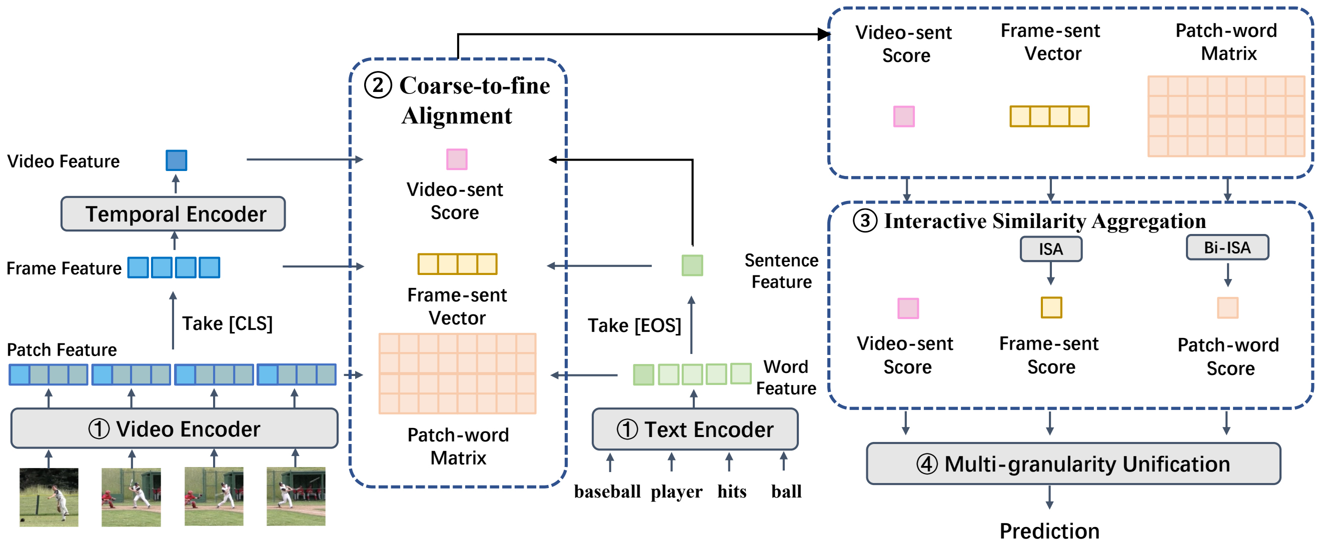

Figure 2. The overview of UCoFiA. It consists of four components: (1) text and video encoders, (2) the coarse-to-fine alignment module which consists of video-sentence alignment (top middle), frame-sentence alignment (center), and patch-word alignment (bottom middle), (3) the interactive similarity aggregation module (including ISA for frame-sentence vector and Bidirectional-ISA (Bi-ISA) for patch-word matrix), and (4) the multi-granularity unification module to aggregate the similarity score from different granularity. (For simplicity, "video-sent" stands for video-sentence, and frame-sent stands for frame-sentence).

and the sentence feature $s \in \mathbb{R}^{C}$ and get the video-sentence similarity score s_{VS}.

Frame-sentence Alignment. To obtain the cross-modal similarity between frames and the sentence, we separately compute the cosine similarity between each row in the frame feature $f \in \mathbb{R}^{N\times C}$ and the sentence feature $s \in \mathbb{R}^{C}$ and get the frame-sentence similarity vector $\mathbf{c}_{\text{FS}} \in \mathbb{R}^{N}$.

Patch-word Alignment. We then consider the fine-grained alignment between text and video. Since the high-level features fail to convey detailed cross-modal information, utilizing the low-level features helps the model to capture subtle correspondence between text and video. To match the feature granularity, we utilize patch and word features for fine-grained alignment. However, due to the high redundancy of patch features, it is ineffective to align all patches to words. Inspired by [35], we adopt an MLP-based patch selection module $\mathcal{H}$ to select the top-K salient patches from each frame according to the frame and video feature that correspond to the patch. The selected patch feature can be denoted as $\hat{p} = \mathcal{H}(p, f, v) \in \mathbb{R}^{L_v\times C}$, where $L_v = N * K$. Afterward, by computing the cosine similarities between all combinations in rows in $\hat{p}$ and rows in the word feature $w \in \mathbb{R}^{L_t\times C}$, we obtain the patch-word similarity matrix $\mathbf{C}_{\text{PW}} \in \mathbb{R}^{L_v\times L_t}$.

Overall, the coarse-to-fine alignment module allows our model to capture cross-modal similarity from the different granularity of features. In Table 3, we demonstrate the effectiveness of each level alignment quantitatively.

3.3. Interactive Similarity Aggregation

Next, we describe how we aggregate the cross-modal similarity vector and matrix for different alignments. Due to the high redundancy of video features, irrelevant information within the similarity vector could impede the model from learning precise cross-modal correspondence. Existing methods [35] propose a softmax-based weighted combination of the similarity vector to reduce the impact of irrelevant information. However, the softmax weights fail to capture the information between input features. For instance, while aggregating the frame-sentence alignment, softmax weights ignore the temporal information across frames. As a result, the weighted combination only focuses on the cross-modal relevance between text query and video frames and ignores the interaction between different frames. To this end, we propose a simple, yet effective interactive similarity aggregation module (ISA).

The idea of ISA module is to jointly consider the cross-modal relevance and the interaction between different features while calculating the weights of different similarities. Inspired by [35, 40], we first compute the cross-modal relevance by applying a softmax layer on the similarity vector. We then adopt a linear layer $\mathcal{T}_i$ on the cross-modal relevance to encourage the interaction between different features. Finally, we apply another softmax layer to obtain the final weight of the similarity vector. ISA module enables the model to better focus on the related features while aggregating the similarity vector. Specifically, given frame-

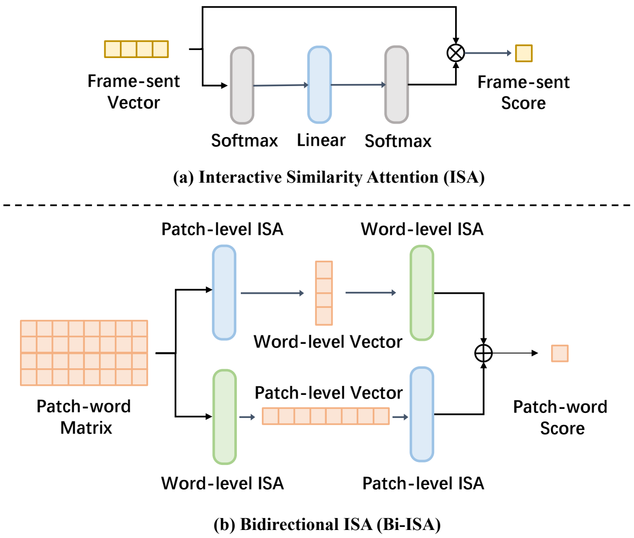

Figure 3. Interactive Similarity Aggregation module (ISA). (a) We directly leverage the ISA module to aggregate the frame-sentence vector to obtain the frame-sentence score. (b) Further, we extend the ISA module to bidirectional ISA (Bi-ISA) to aggregate the patch-word matrix. (For simplicity, frame-sent stands for frame-sentence).

sentence similarity vector $\mathbf{c}_{\text{FS}} \in \mathbb{R}^N$, the ISA module can be formulated as:

$$s_{\text{FS}} = \text{Softmax}\left(\mathcal{T}_i\left(\text{Softmax}\left(\mathbf{c}_{\text{FS}}\right)\right)\right)\mathbf{c}_{\text{FS}}, \tag{1}$$

where s_{FS} denotes the frame-sentence similarity score between the text query and video candidate. To sum up, ISA module is capable of aggregating the similarity vector to obtain a similarity score by jointly considering cross-modal relevance and feature interaction. Regardless of the feature dimension, ISA module is flexible to deal with similarity vectors with different feature granularity. Thus, we extend the ISA module to aggregate the patch-word similarity matrix.

A direct idea is to flatten the patch-word matrix to a large vector and apply the ISA module to obtain the similarity score. However, due to the modality gap, it is difficult to model the feature interaction across video and text for ISA module. The alternative idea is to split each row or column of the similarity matrix into a similarity vector and leverage the ISA module on each vector. Afterward, we aggregate the similarity score from each vector to another similarity vector and apply another ISA to obtain the patch-word score. In that way, we can separately model the feature interaction between patches and words to provide better similarity aggregation. We consider two ways of aggregation: patch-then-word (see the top of Figure 3(b)) and word-then-patch (see the bottom of Figure 3(b)). Empirically we find that jointly considering these two directions provide better aggregation for the patch-word matrix. To this end, we combine the strength of both directions and name the module as a bi-directional ISA module (Bi-ISA, Figure 3(b)). To describe the module formally, we first adopt a patch-level ISA module $\mathcal{A}_p$ on $\mathbf{C}_{\text{PW}}$ to obtain a word-level similarity vector. We then adopt a word-level ISA module $\mathcal{A}_w$ to aggregate the word-level similarity vector to the patch-then-word score. Similarly, we can obtain the word-then-patch score by leveraging a word-level ISA module and a patch-level ISA in a reverse way. The whole process of Bi-ISA can be formulated as:

$$s_{\text{PW}} = \mathcal{A}_p\left(\mathcal{A}_w\left(\mathbf{C}_{\text{PW}}\right)\right) + \mathcal{A}_w\left(\mathcal{A}_p\left(\mathbf{C}_{\text{PW}}\right)\right), \tag{2}$$

where s_{PW} denotes the patch-word similarity score. Conceptually, our ISA module jointly considers the cross-modal relevance and the interaction between different features while aggregating the similarity vector (matrix). In Table 5 and Table 6, we validate the effectiveness of our ISA and Bi-ISA module compared to different aggregation mechanisms. In the next section, we further aggregate the different levels of similarity score to one final score for retrieval.

3.4. Unifying Coarse and Fine-grained Alignments

For simplicity, we explained our approach above with only one video-query pair; but when we perform retrieval on G videos and H queries, we compute the similarity scores over all possible combinations of the videos and queries, and denote the score combinations for each alignment level (video-sentence, frame-sentence, patch-word) as $\mathbf{S}_{\text{VS}}, \mathbf{S}_{\text{FS}}, \mathbf{S}_{\text{PW}} \in \mathbb{R}^{G \times H}$, where $\mathbf{S}^{ij}$ is the score for the i^{th} video and j^{th} query. For the last step of obtaining the final cross-modal similarity for retrieval, the common method [40] is usually to directly compute the average over the different levels of similarities.

However, we find that scores across different videos are highly imbalanced in the similarity matrices of each level, and we empirically find that correcting the issue before summing the similarities leads to a better result in multi-level alignment methods [40, 35]. The imbalance issue is similar to the findings of Park *et al.* [42]. Specifically, sometimes the summation of retrieval similarities between one specific video and all texts (we called this term *marginal similarity* in the following) might be much higher than that of the other videos, meaning that this video is over-represented and will lower the probability of the other video being selected. To address this, we re-scale the similarity matrix to normalize the marginal similarity of every video to be a similar value. One approach is to apply dual softmax operation [11] on the similarity matrix, but this is not realistic in the testing phase since it requires obtaining all the testing videos and queries at hand.

A more realistic setting is where we can access all the testing videos, but only have one test query at a time for text-to-video retrieval(reversely, for video-to-text retrieval, we search the best candidate out of all testing queries for one test video). Since there is only one data point on the query side, double softmax is not useful in this case. To address this, we use the training query set as the approximation for the test set and follow Park *et al*. [42] to leverage the Sinkhorn-Knopp algorithm [14] to correct the imbalance. We describe our algorithm for text-to-video retrieval for simplicity as we can just swap videos and texts in the algorithm for video-to-text retrieval. Specifically, assuming we have G test videos and J train queries, for each level of similarity score, the algorithm computes the test video bias $\alpha \in \mathbb{R}^G$ and training text bias $\beta \in \mathbb{R}^J$ in an alternating and iterative manner. Please refer to the supplements for the detail of the algorithm. Adding these biases to the similarity matrix can normalize it to have similar marginal similarities for videos and for texts, respectively. However, since we only introduce the training query set to approximate the test set distribution and help to compute the test video bias α, we discard the training query bias β after. We then add α to the testing similarity matrix $\mathbf{S} \in \mathbb{R}^{G\times H}$ that is generated by G test videos and H test queries, that is,

$$\mathrm{SK}(\mathbf{S})^{ij} = \mathbf{S}^{ij} + \alpha^i \tag{3}$$

We then obtain a normalized similarity matrix for testing videos. We apply the algorithm separately on the similarity matrix of different alignments before summing them together, and we empirically find out this is better than doing summation first and then normalization. Finally, the final retrieval score $\mathbf{R}$ can be written as:

$$\mathbf{R} = \mathrm{SK}(\mathbf{S}_{\mathrm{VQ}}) + \mathrm{SK}(\mathbf{S}_{\mathrm{SQ}}) + \mathrm{SK}(\mathbf{S}_{\mathrm{PQ}}) \tag{4}$$

Note that we only apply Equations (3) and (4) in the inference phase. Similarly, for video-to-text retrieval, we normalize the similarity matrix by adding the test text bias. In Table 7, we validate the effectiveness of applying the above normalization. We also provide a visualization of the reduction of over-/under-representation by applying this technique in the supplementary material.

3.5. Training and Inference

During training, we randomly sample B video-query pairs, compute the scores over all possible combinations between the videos and the queries without normalization, and denote the similarity combination as $\mathbf{R} \in \mathbb{R}^{B\times B}$, where $\mathbf{R}^{ij}$ is the score for the i^{th} video and j^{th} query. We utilize the cross-modal contrastive objective [44] to maximize the scores of the positive pairs (the diagonal in $\mathbf{R}$) and minimize the scores of negative pairs:

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp\left(\mathbf{R}^{ii}\right)}{\sum_{j=1}^{B}\exp\left(\mathbf{R}^{ij}\right)}, \tag{5}$$

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp\left(\mathbf{R}^{ii}\right)}{\sum_{j=1}^{B}\exp\left(\mathbf{R}^{ji}\right)}, \tag{6}$$

$$\mathcal{L} = \mathcal{L}_{v2t} + \mathcal{L}_{t2v}, \tag{7}$$

During inference, to perform video-text retrieval, we will compute the similarities between all videos and the query, normalize the similarities by the method introduced in Section 3.4, and retrieve the video with the highest similarity. We conduct the procedure similarly but in the other direction for video-to-text retrieval.

4. Experimental Setup

4.1. Datasets

We evaluate UCoFiA on five popular video-text retrieval datasets: MSR-VTT [57], MSVD [7], ActivityNet [25] and DiDeMo [2].

MSR-VTT [57] contains $10,000$ videos, each annotated with 20 text captions. The video length is ranged from 10 to 32 seconds. Following [34, 19], we train UCoFiA on $9,000$ videos and report the results on $1,000$ selected video-text pairs (the 1kA test set).

Activity-Net [25] consists of $20,000$ YouTube videos with $100,000$ captions. The average video length is 180 seconds. We follow [39] to concatenate the multiple text descriptions of a video into one paragraph and perform paragraph-to-video retrieval on ActivityNet. We train our model on $10,000$ videos and use the 'val1' split for evaluation which contains $5,000$ videos.

DiDeMo [2] is comprised of $10,000$ videos and $40,000$ captions. The average video length is 30 seconds. Similar to ActivityNet, we evaluate paragraph-to-video retrieval on DiDeMo. There are $8,395$ videos in the training set, $1,065$ videos in the validation set, and $1,004$ videos in the test set. We report the results on the test set for evaluation.

MSVD [7] contains $1,970$ videos, each with a length that ranges from 1 to 62 seconds. Each video contains approximately 40 captions. Following [39], we split the train, validation, and test set with $1,200$, 100, and 670 videos. We follow the common multiple caption evaluation [34, 39] setting in which each video in the test set is associated with multiple text captions.

VATEX [52] consists of $34,991$ video clips with multiple captions per video. We follow HGR's [8] split protocol. There are $25,991$ videos in the training set, $1,500$ videos in the validation set and $1,500$ videos for evaluation.

4.2. Evaluation Metrics

Following [39], we use standard video-text retrieval metrics, including R@1, R@5, and Mean Rank (MnR) to validate the effectiveness of our UCoFiA model. We report re-

Method	MSR-VTT			Activity-Net			DiDeMo			MSVD			VATEX		
	R@1	R@5	MnR↓	R@1	R@5	MnR↓	R@1	R@5	MnR↓	R@1	R@5	MnR↓	R@1	R@5	MnR↓
Non-CLIP Methods															
CE [34]	20.9	48.8	28.2	18.2	47.7	23.1	16.1	41.1	43.7	19.8	49.0	-	-	-	-
MMT [19]	26.6	57.1	24.0	26.6	57.1	16.0	-	-	-	-	-	-	-	-	-
Support Set [43]	30.1	58.5	-	29.2	61.6	-	-	-	-	28.4	60.0	-	45.9	82.4	-
Frozen [3]	31.0	59.5	-	-	-	-	34.6	65.0	-	33.7	64.7	-	-	-	-
All-in-one [51]	37.9	68.1	-	22.4	53.7	-	32.7	61.4	-	-	-	-	-	-	-
Singularity [27]	*42.7*	*69.5*	*-*	*48.9*	*77.0*	*-*	*53.1*	*79.9*	*-*	*-*	*-*	*-*	*-*	*-*	*-*
VindLU [10]	*45.3*	*69.9*	*-*	*54.4*	*80.7*	*-*	*59.2*	*84.1*	*-*	*-*	*-*	*-*	*-*	*-*	*-*
CLIP-based Methods															
CLIP4Clip [39]	44.5	71.4	15.3	40.5	73.4	10.0	43.4	70.2	17.5	46.2	76.1	10.0	55.9	89.2	3.9
CAMoE [11]	44.6	72.6	13.3	-	-	-	-	-	-	46.9	76.1	9.8	-	-	-
X-Pool [20]	46.9	72.8	14.3	-	-	-	-	-	-	47.2	77.4	**9.3**	-	-	-
TS2-Net [35]	47.0	**74.5**	13.0	41.0	73.6	8.4	41.8	71.6	14.8	-	-	-	59.1	90.0	3.5
X-CLIP [40]	46.1	73.0	13.2	44.3	74.1	7.9	45.2	74.0	14.6	47.1	**77.8**	9.5	-	-	-
UCoFiA	**49.4**	72.1	**12.9**	**45.7**	**76.0**	**6.6**	**46.5**	**74.8**	**13.4**	**47.4**	77.6	9.6	**61.1**	**90.5**	**3.4**

Table 1. Comparison to the state-of-the-art text-to-video retrieval methods on MSR-VTT, AcitivityNet, DiDeMo, MSVD, VATEX. The top section shows the results of non-CLIP methods and the middle section shows the results of CLIP-based methods. Our results indicate that UCoFiA achieves better or comparable results on all five datasets compared to the current state-of-the-art methods. For a fair comparison, we de-emphasize Singularity [27] and VindLU [10] (by using gray color and italic font) since they are pretrained on large-scale video datasets and use time-consuming two-stage re-ranking strategy(two-step retrieval, the first step retrieves top K candidates from all candidates and the second step retrieves from the top K candidates).

Method	MSR-VTT		Activity-Net		DiDeMo	
	R@1	R@5	R@1	R@5	R@1	R@5
CE [34]	20.6	50.3	17.7	46.6	15.6	40.9
MMT [19]	27.0	57.5	28.9	61.1	-	-
Support set [43]	26.6	55.1	28.7	60.8	-	-
HiT [33]	32.1	62.7	-	-	-	-
TT-CE [13]	32.1	62.7	23.0	56.1	21.1	47.3
CLIP4Clip [39]	42.7	70.9	42.5	74.1	42.5	70.6
TS2-Net [35]	45.3	74.1	-	-	-	-
X-CLIP [40]	46.8	73.3	43.9	73.9	43.1	**72.2**
UCoFiA	**47.1**	**74.3**	**46.3**	**76.5**	**46.0**	71.9

Table 2. Comparison to the state-of-the-art video-to-text retrieval methods on MSR-VTT, AcitivityNet, DiDeMo datasets. The top section shows the results of non-CLIP methods and the middle section shows the results of CLIP-based methods. Our results indicate that UCoFiA achieves better or comparable results on all datasets compared to the current state-of-the-art methods.

sults with more metrics (including R@10, Median Recall) in supplements.

4.3. Implementation Details

Following [39], we leverage the CLIP [44] pre-trained weights to initialize our UCoFiA model. For the visual encoder, we use CLIP's ViT-B/32 weights. The dimension C of visual and textual representations is set to 512. We choose the $K = 4$ most salient patches for each frame in our patch selection module on all datasets. For MSR-VTT, MSVD, and VATEX, we follow the previous work [39] to sample $N = 12$ frames per video and set the max length of text query as 32. For the paragraph-to-video dataset ActivityNet and DiDeMo, we sample 64 frames per video and set the max length of text query as 64. We adopt the Adam optimizer [24] with a cosine warm-up strategy [36]. We set the learning rate of the visual and text encoders as $1e{-7}$ and other modules as $1e{-4}$. In the Sinkhorn-Knopp algorithm, we set the number of iterations as 4 across all datasets. We set a batch size to 128 for MSR-VTT, MSVD, and VATEX and 64 for ActivityNet and Didemo following [40]. We train UCoFiA for $8, 5, 20, 20, 20$ epochs on MSR-VTT, MSVD, ActivityNet, DiDeMo, and VATEX, respectively. We conduct the ablation study on the most popular MSR-VTT dataset to analyze the effect of different designs of our model.

5. Experimental Results

In this section, we compare UCoFiA with several recent methods on the five video-text retrieval datasets and conduct comprehensive ablation studies to verify our design choices. We also provide a qualitative analysis to show the effectiveness of our model designs. Moreover, we display quantitative results with full metrics (R@1,5,10, MdR, MnR) for each dataset, more experiments about applying UCoFiA to more advanced backbone model [58], and quantitative analysis on computational cost and training strategy in the supplements.

5.1. Comparison to State-of-the-art Approaches

In Table 1, we compare UCoFiA with existing methods that are either with (in the middle section of the table) or without (in the upper section of the table) CLIP on text-to-video retrieval. We also compare UCoFiA with existing methods on video-to-text retrieval in Table 2. We observe that UCoFiA achieves better performance than existing methods on most of the metrics on both text-to-video and video-to-text retrieval settings.

On MSR-VTT, compared to the recent multi-level alignment method X-CLIP [40], UCoFiA gives a significant 3.3% improvement on text-to-video R@1 metric and comparable video-to-text retrieval results despite X-CLIP leveraging more levels of coarse alignments (video-sentence, video-word, frame-sentence, and frame-word) than us. This result verifies our motivation that building a coarse-to-fine alignment is useful for video-text retrieval. We also achieve 2.0% improvement on text-to-video R@1 on VATEX dataset.

Similarly, on ActivityNet and DiDeMo datasets, we notice that UCoFiA is capable of handling longer text queries and achieves observe 5.2% and 4.7% improvement on paragraph-to-video retrieval compared to CLIP4Clip [39] which only utilizes video-sentence alignment. Furthermore, we observe 1.4% and 1.3% gain on ActivityNet and DiDeMo compared to X-CLIP [40] under paragraph-to-video retrieval and 2.4% and 2.9% gain on video-to-paragraph retrieval. These results show the importance of fine-grained correspondence even on retrieval for long videos.

Meanwhile, our method achieves comparable results on the MSVD dataset which evaluates a multiple-caption setting, which further verifies the generalizability of our model. However, we observe our normalization before the summation strategy still introduces some performance gain on MSVD even though the multiple-caption setting breaks our assumption that one video has one corresponding query, showing the robustness of our approach.

5.2. Ablation study

In this section, we study the different design choices of our UCoFiA model and verify their effects on the video-text retrieval performance on MSR-VTT under text-to-video retrieval setting. Specifically, we investigate (1) the effect of different alignment schemes, (2) the comparison of our fine-grained alignment design to others, (3) different similarity aggregation methods and (4) the effect of the unification module.

The Effect of Different Alignment Schemes. First, we validate the effectiveness of our different levels of alignment. As shown in Table 3, adding frame-sentence and patch-word alignments improves the base model (that only leverages video-sentence alignment) with a significant margin. Specifically, we observe that adding patch-word alignment improves the R@1 while keeping a similar R@5, which indicates that the fine-grained alignment helps the model choose the most relevant video (top-1) from several similar video candidates (top-5) by capturing the subtle differences between these video candidates. This result justifies our motivation for using fine-grained alignment as complementary to coarse alignment.

video-sent	frame-sent	patch-word	R@1	R@5	MnR↓
✓			44.5	71.1	15.2
✓	✓		47.1	73.2	14.1
✓	✓	✓	**48.2**	**73.3**	**13.2**

Table 3. The effect of different level alignments. Video-sent denotes the video-sentence alignment, frame-sent denotes the frame-sentence alignment and patch-word denotes patch-word alignment. The results indicate the effectiveness of each level alignment.

patch-sent	patch-word	R@1	R@5	MnR↓
		47.1	73.2	14.1
✓		46.5	**73.7**	14.7
	✓	**48.2**	73.3	**13.2**
✓	✓	47.2	71.9	13.8

Table 4. Comparison of our fine-grained alignment design to others. "Patch-word" denotes patch-word alignment and "patch-sentence" denotes patch-sentence alignment. The last row denotes the ensemble of patch-word and patch-sentence alignment. The results indicate our patch-word alignment is the best design choice for fine-grained alignment on MSR-VTT.

Aggregation Method	R@1	R@5	MnR ↓
Mean Pooling	45.6	71.2	15.2
Softmax Weight	46.2	72.7	14.7
ISA (ours)	**47.1**	**73.2**	**14.1**

Table 5. Effect of different similarity aggregation methods for frame-sentence vector. Results show that our ISA module has better performance on all metrics.

Comparison of Our Fine-grained Alignment Design to Others. In Table 4, we study the different designs of fine-grained alignment in our model. We compare our patch-word similarity with patch-sentence alignment and the ensemble of these two alignments. Based on the results, we observe that our patch-word alignment is the best design, and meanwhile, adding patch-sentence alignment degrades the performance, possibly caused by the mismatch between patch and sentence representations, where one conveys local information and the other contains global information. We also provide qualitative analysis on different alignment designs in supplements.

Different Similarity Aggregation Methods. To validate the effectiveness of our ISA module for frame-sentence alignment and Bi-ISA module for patch-word alignment in Section 3.3, we compare our modules with several other ag-

Aggregation Method	Bidirectional	R@1	R@5	MnR↓
Direct-ISA		46.4	72.8	14.8
Mean Pooling	✓	47.1	73.0	13.9
Softmax Weight	✓	47.5	**73.4**	13.5
Bi-ISA (ours)	✓	**48.2**	73.3	**13.2**

Table 6. Effect of different similarity aggregation methods for patch-word matrix. Results verify the effectiveness of our Bi-ISA module compared to other methods.

Setting	R@1	R@5	R@10	MnR↓
w/o SK norm	48.2	**73.3**	82.3	13.2
w SK norm	**49.4**	72.1	**83.5**	**12.9**

Table 7. The effect of SK norm of different level similarity scores.

gregation methods. In Table 5, we show the effectiveness of our interactive similarity attention (ISA) on the frame-sentence score. Specifically, we compare the vanilla mean pooling strategy and softmax-based weighted combination [35]. Note that we only adopt video-sentence and frame-sentence alignment (remove patch-word alignment and the Bi-ISA module) for the experiments in Table 5 to study the effect of ISA independently. Results show that using the ISA module achieves better performance compared to other aggregation methods. We also compare our Bi-ISA with other aggregation methods in Table 6, where we have adopted ISA for the frame-sentence score. We observe that the design of bi-directional aggregation achieves a significant gain in performance. In general, our experiments show that adding one linear layer to learn the temporal information across frames before aggregation is crucial.

The Effect of Unification Module. To verify the importance of the unification module for different levels of similarity, we compare our UCoFiA model with the variant that removes the Sinkhorn-Knopp normalization (abbreviated as SK norm) in Table 7. We notice that adding the SK norm provides better performance on video-text retrieval with a 1.2% gain on both R@1 and R@10.

5.3. Qualitative Analysis

In this section, we report the qualitative analysis on the effectiveness of the ISA module. We show the comparison of the model with and without ISA module in Figure 4. We can see from the upper part of Figure 4 that the ISA module improves the softmax weight by highlighting the most relevant frame to the query. As a result, the calculated similarity score of ISA is significantly increased since the irrelevant information is eliminated. In the lower part, by recognizing the most related frame, our model with ISA produces a lower similarity to the unmatched video. For either the ground truth or wrongly retrieved video, we find the softmax weights used in TS2-Net[35] (without frame-wise interaction) tend to be more uniformly distributed. On the contrary, our ISA module is capable of finding and assigning the highest score to the most relevant frame (the player is making shots) through the interaction between frames, which shows video candidates and results in the correct retrieval. This analysis verifies the effectiveness of our ISA module.

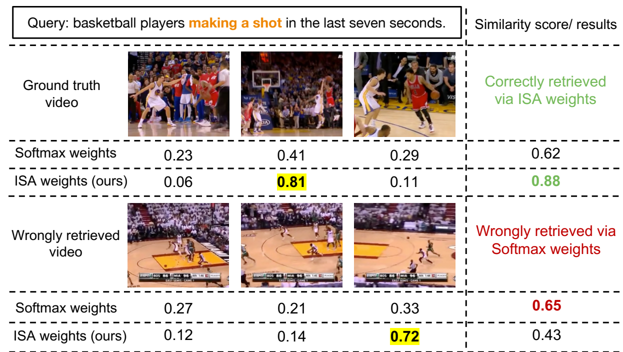

Figure 4. The visualization of the effectiveness of our ISA module. In the upper part, we show that the ISA improves the softmax weight by highlighting the most relevant frame to the query. As a result, the calculated similarity score of ISA is significantly increased since the irrelevant information is eliminated. In the lower part, we show that the ISA module is also capable of assigning the most related frame (the player is about to get the ball) the highest score for the unmatched video (caption: the player is chasing the ball). However, since it is still not directly related to the text query (compared to the upper video), thus our model with ISA produces a lower similarity score.

6. Conclusion

In this paper, we present UCoFiA, which jointly considers cross-modal correspondence from different granularity and accomplishes the unification of multi-grained alignment. It achieves state-of-the-art results on multiple video-text retrieval benchmarks. UCoFiA is a simple but effective model that achieves state-of-the-art results on five diverse video-text retrieval benchmarks. In the future, we plan to extend our method to other video-language tasks such as video question answering and video reasoning.

Acknowledgment

We thank the reviewers and Shoubin Yu for helpful discussions. This work was supported by ARO Award W911NF2110220, ONR Grant N00014-23-1-2356, DARPA KAIROS Grant FA8750-19-2-1004, NSF-AI Engage Institute DRL211263, Sony Faculty Innovation award, and Laboratory for Analytic Sciences via NC State University.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch,

Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 2

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 6

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2, 7

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker's guide to long video retrieval. *arXiv preprint arXiv:2205.08508*, 2022. 2

[5] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5205, 2022. 3

[6] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927, 2022. 2

[7] David Chen and William B Dolan. 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, 2011. 6

[8] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 2, 6

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 2

[10] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. *arXiv preprint arXiv:2212.05051*, 2022. 2, 7

[11] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 1, 3, 5, 7

[12] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19016–19026, 2023. 2

[13] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 2, 7

[14] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 2, 6

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[16] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 2

[17] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3363, 2021. 2

[18] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 1, 2

[19] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020. 6, 7

[20] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5006–5015, 2022. 1, 2, 7

[21] Yunhao Gou, Tom Ko, Hansi Yang, James Kwok, Yu Zhang, and Mingxuan Wang. Leveraging per image-token consistency for vision-language pre-training. *arXiv preprint arXiv:2211.15398*, 2022. 2

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[23] Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*, 2022. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[25] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In

Proceedings of the IEEE international conference on computer vision, pages 706–715, 2017. 6

[26] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 2

[27] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 2, 7

[28] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1, 2

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[30] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022. 2

[31] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 413–430. Springer, 2022. 2

[32] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309, 2023. 2

[33] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11915–11925, 2021. 7

[34] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 6, 7

[35] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. 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*, pages 319–335. Springer, 2022. 1, 2, 3, 4, 5, 7, 9

[36] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[37] Haoyu Lu, Mingyu Ding, Nanyi Fei, Yuqi Huo, and Zhiwu Lu. Lgdn: Language-guided denoising network for video-language modeling. *arXiv preprint arXiv:2209.11388*, 2022. 2

[38] Haoyu Lu, Mingyu Ding, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Masayoshi Tomizuka, and Wei Zhan. Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. *arXiv preprint arXiv:2302.06605*, 2023. 2

[39] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 1, 2, 3, 6, 7, 8

[40] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 1, 3, 4, 5, 7, 8

[41] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23, 2021. 2

[42] Yookoon Park, Mahmoud Azab, Bo Xiong, Seungwhan Moon, Florian Metze, Gourab Kundu, and Kirmani Ahmed. Normalized contrastive learning for text-video retrieval. *arXiv preprint arXiv:2212.11790*, 2022. 3, 5, 6

[43] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 7

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 7

[45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[46] Fangxun Shu, Biaolong Chen, Yue Liao, Shuwen Xiao, Wenyu Sun, Xiaobo Li, Yousong Zhu, Jinqiao Wang, and Si Liu. Masked contrastive pre-training for efficient video-text retrieval. *arXiv preprint arXiv:2212.00986*, 2022. 2

[47] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005, 2022. 1

[48] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 1

[49] Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. An empirical study of multimodal model merging. *arXiv preprint arXiv:2304.14933*, 2023. 1

[50] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016. 1, 2

[51] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Explor-

ing unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 7

[52] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 6

[53] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2021. 2

[54] Yimu Wang and Peng Shi. Video-text retrieval by supervised multi-space multi-grained alignment. *arXiv preprint arXiv:2302.09473*, 2023. 2

[55] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *arXiv preprint arXiv:2205.10747*, 2022. 2

[56] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 1

[57] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 6

[58] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 7

[59] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. 2

[60] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[61] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023. 1

[62] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 1, 2

[63] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. Video captioning and retrieval models with semantic attention. *arXiv preprint arXiv:1610.02947*, 6(7), 2016. 1, 2

[64] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 1, 2

[65] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 2

[66] Xiaohan Zou, Changqiao Wu, Lele Cheng, and Zhongyuan Wang. Tokenflow: Rethinking fine-grained cross-modal alignment in vision-language retrieval. *arXiv preprint arXiv:2209.13822*, 2022. 2, 3