Dig into Multi-modal Cues for Video Retrieval with Hierarchical Alignment

Wenzhe Wang[1] , **Mengdan Zhang**[2] , **Runnan Chen**[3] , **Guanyu Cai**[4] , **Penghao Zhou**[2] , **Pai Peng**[2] , **Xiaowei Guo**[2] , **Jian Wu**[1,*] , **Xing Sun**[2,*]
[1]Zhejiang University, China
[2]Youtu Lab, Tencent, China
[3]The University of Hong Kong, China
[4]Tongji University, China
{wangwnezhe, wujian2000}@zju.edu.cn, rnchen2@cs.hku.hk, caiguanyu@tongji.edu.cn,
{davinazhang, penghaozhou, popeyepeng, scorpioguo, winfredsun}@tencent.com

Abstract

Multi-modal cues presented in videos are usually beneficial for the challenging video-text retrieval task on internet-scale datasets. Recent video retrieval methods take advantage of multi-modal cues by aggregating them to holistic high-level semantics for matching with text representations in a global view. In contrast to this global alignment, the local alignment of detailed semantics encoded within both multi-modal cues and distinct phrases is still not well conducted. Thus, in this paper, we leverage the hierarchical video-text alignment to fully explore the detailed diverse characteristics in multi-modal cues for fine-grained alignment with local semantics from phrases, as well as to capture a high-level semantic correspondence. Specifically, multi-step attention is learned for progressively comprehensive local alignment and a holistic transformer is utilized to summarize multi-modal cues for global alignment. With hierarchical alignment, our model outperforms state-of-the-art methods on three public video retrieval datasets.

1 Introduction

Vision and language play important roles in the way humans learn to associate visual entities to abstract concepts and vice versa. With the rapid emergence of videos on the Internet, video-text retrieval has become challenging, since both videos and language texts contain rich and structured details.

As shown in Figure 1, a single piece of text is able to demonstrate complicated interactions among various entities, where actions (e.g. 'playing' and 'standing') are denoted by verbs and entities refer to noun phrases (e.g. 'horse' and 'bucket'). Meanwhile, these complicated cues can be well described by constituent modalities of the video data including appearance, motion, audio, overlaid text, speech, etc. Therefore, more and more recent works focus on taking advantage of multi-modal video information for accurate video-text retrieval. For example, benefiting from multi-modal features

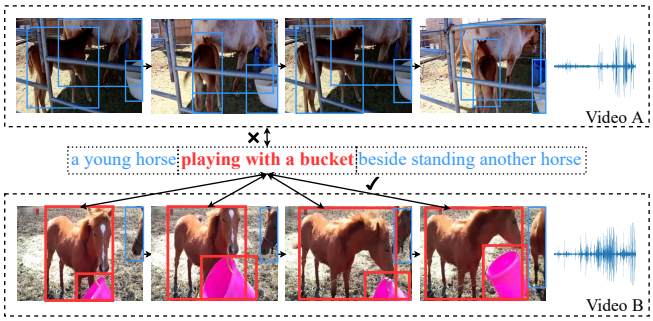

Figure 1: An example of text-to-video retrieval. High-level semantics can be summarized from video cues of 'young horse', 'playing', 'bucket', 'standing', 'another horse', etc. Methods only aligning high-level semantics may not know whether 'young horse beside standing another horse' or 'young horse playing with a bucket' is the key to distinguish the videos, since the former one is also distinctive among the retrieval videos in a large dataset. In this paper, hierarchical alignment is utilized to help mitigate this issue.

extracted by multiple off-the-shelf expert models, compact global video representations are learned in [Liu *et al.*, 2019; Gabeur *et al.*, 2020; Patrick *et al.*, 2020]. These global video representations aggregate high-level semantics and are able to be well aligned to the corresponding encoded global text representations. These methods usually achieve promising results for video-text pairs containing distinct semantics. However, miss matching may occur for these methods when video-text pairs contain similar dominant semantics. It is because global alignment among high-level semantics may overlook important details or local relationships contained in local phrases or specific video modalities.

To demonstrate the above issue more concretely, we give an example of text-to-video retrieval in Figure 1 where video A and B contain similar dominant semantics. The two videos both contain modality cues of 'young horse', 'playing', 'bucket', etc, as mentioned in the text. Global-alignment-based methods learn to summarize all these cues and their relationships into a consistent and compact embedding. But they may not know whether 'young horse beside standing another horse' or 'young horse playing with a bucket' is the key to distinguish the two videos, since the former one is

* Corresponding authors.

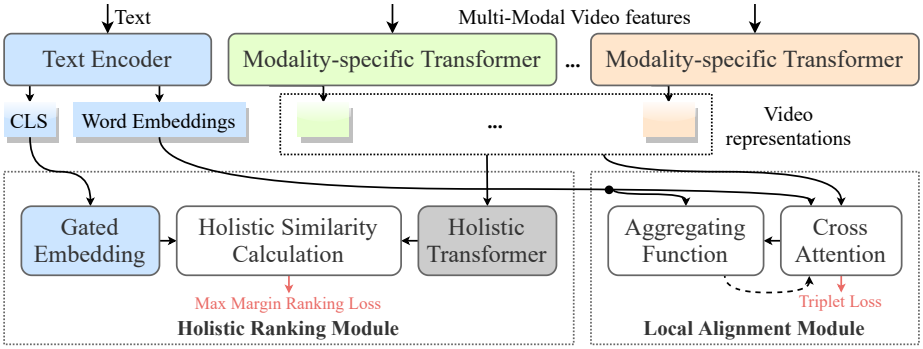

Figure 2: Illustration of our model for text and video bi-directional retrieval. Arrows in black denote the data flow direction, those in red denote loss function, and those in dotted denote multi-step path. Best view in color.

also distinctive among the retrieval videos in a large dataset. This suggests the importance of introducing local alignment. However, if only aligning each word embedding in texts with each modality in videos and ignoring relatively higher-level (e.g., phrases in texts) semantics, both video A and B can be the retrieval result. Therefore, it is essential for video-text retrieval to capture detailed correspondences and meanwhile avoid being trapped in local optimum matching.

In this paper, we propose to dig into multi-modal cues for video-text bi-directional retrieval with hierarchical alignment. For the multi-modal cues, our model learns detailed modality-specific semantics and their local relationships, and also the global induction of high-level semantics. Hierarchical alignment is carried out to align all the local semantics and relations, and meanwhile to ensure a high-level correspondence. Specifically, to extract detailed semantics, several modality-specific transformers are utilized to capture intrinsic characteristics of each modality such as the temporal and spatial contextual properties. Then, local alignment is achieved by attending modality-specific video representations to the embedding of each word by multi-step cross attention. The importance of each word in a sentence is thus learned, and the relations of the words and diverse multi-modal cues are captured. Moreover, a holistic transformer is used for aggregating high-level semantics (e.g., descriptions of various events in videos) into the global video representation. The representation is then used for global alignment to the global text embeddings and for retrieval ranking.

Extensive experiments are conducted on the MSR-VTT [Xu *et al.*, 2016], ActivityNet [Caba Heilbron *et al.*, 2015], and LSMDC [Torabi *et al.*, 2016] to evaluate video-text bi-directional retrieval performance. Experimental results show that our proposed model can achieve state-of-the-art results on all the above datasets. Ablation studies are carried out to evaluate the effectiveness of each part of our model.

The contributions of our work are two folds. 1) We propose a model that leverages the hierarchical video-text alignment to fully explore the detailed and diverse characteristics in multi-modal cues for fine-grained alignment with local textual semantics, as well as to capture a high-level semantic correspondence. 2) Our proposed model can achieve new state-of-the-art results on three benchmark datasets, demonstrating its effectiveness and generalization.

2 Related Works

2.1 Self-supervised Learning of Video

[Luo *et al.*, 2020; Miech *et al.*, 2020; Rouditchenko *et al.*, 2020; Zhu and Yang, 2020] extracted text or audio information in videos and used it for self-supervised learning on large datasets like HowTo100M [Miech *et al.*, 2019]. [Luo *et al.*, 2020; Zhu and Yang, 2020] introduced new Transformer blocks for textual and visual feature encoding and learning. [Miech *et al.*, 2020] proposed an approach named MIL-NCE to learn strong video representations from scratch. And [Rouditchenko *et al.*, 2020] proposed a network that learns audio-visual language representations directly from randomly segmented video clips and their raw audio waveform.

2.2 Text-Video Retrieval

For better feature matching in the text-video retrieval task, [Chen *et al.*, 2020a; Wei *et al.*, 2020] focused on metric learning methods that assigning weights to positive and negative pairs respectively during network training. An encoding architecture based entirely on convolutional neural networks is proposed by [Li *et al.*, 2020] for better text and video feature encoding. And [Zhao *et al.*, 2020] focused on the long-range dependency in videos and texts by introducing a multi-scale dilated convolutional block.

Several works explored the relationships within texts and videos.[Chen *et al.*, 2020c; Yang *et al.*, 2020] built semantic trees for words in texts to match the corresponding information in videos in a coarse-to-fine manner. [Feng *et al.*, 2020; Wang *et al.*, 2020] constructed relation graphs using features from both texts and videos for better aligning similar information during retrieval. Since multi-modal cues are not utilized by these works, their retrieval performance is not promising.

For multi-modal learning, [Liu *et al.*, 2019; Gabeur *et al.*, 2020] utilized pre-trained models from other tasks to extract multi-modal video features for text and video bi-directional retrieval. And [Patrick *et al.*, 2020] further replaced the original retrieval texts with texts generated by a decoder taking video features from support sets as input.

3 Methodology

Our model carries out hierarchical alignment for video retrieval, which takes a text $T = \{w_i\}_{i=1}^{l}$ and a video's multi-modal features $V = \{m_j\}_{j=1}^{n}$ as inputs. We define l as the length of T, w_i as the i-th word, n as the number of modalities in the video, and m_j as the features of the j-th modality.

The main architecture of our model is illustrated in Figure 2, which is composed of four main components, i.e., a text encoder for text feature encoding, several modality-specific transformers [Vaswani *et al.*, 2017] for multi-modal video feature encoding, a Local Alignment Module for multi-modal local information alignment and a Holistic Ranking Module for semantic relationship learning and holistic similarity calculation. In the following of this section, we first introduce how we encode text and video features into representations, then describe the two modules, respectively, and finally define the objective function of our model.

3.1 Local Feature Encoding

For text encoding, a pre-trained Bert model [Devlin *et al.*, 2018] (defined as $\Phi(\cdot)$) is used for fine-tuning, the [CLS] output ($T^c = \Phi(T)^{[c]}$) and word embedding output ($T^w = \{w_i | w_i = \Phi(T)^{(i)}\}_{i=1}^{l}$) of which are fed to the Holistic Ranking Module and Local Alignment Module.

For multi-modal video feature encoding, the multi-modal video features V are acquired by feeding retrieval videos to n video experts pre-trained on other video-related large-scale datasets following [Liu *et al.*, 2019; Gabeur *et al.*, 2020]. Then, the features of each modality are pre-processed and fed to the modality-specific transformers for detailed representation learning. Specifically, all these transformers $\Psi_j^M(\cdot)$ follow the architecture of the encoder of Transformer in [Vaswani *et al.*, 2017], consisting of multi-head attention and fully-connected layers. Since transformers take embeddings as input, we pre-process video features V as follows.

Firstly, since video features V output from different models have different dimensions, we learn n linear layers $\{F_j\}_{j=1}^{n}$ to project the features into a unified dimension d_{uni}:

$$V^{uni} = \left\{ m_j^{uni} | m_j^{uni} = F_j\left(m_j\right) \right\}_{j=1}^{n} \tag{1}$$

For input of the modality-specific transformers, feature and temporal embeddings are calculated. To initialize the feature embeddings, a max-pooling aggregation is utilized for all features in each m_j^{uni}, i.e., $m_j^{agg\text{-}f} = maxpool(m_j^{uni})$, and the feature embeddings of the j-th modality-specific transformer E_j^f is defined as the concatenation of the corresponding aggregated and unified features, i.e., $E_j^f = [m_j^{agg\text{-}f}, m_j^{uni}]$. To use temporal cues in videos, we embed each second's features in m_j^{uni} together as [Gabeur *et al.*, 2020] to achieve m_j^{tmp} and calculate the corresponding aggregated features $m_j^{agg\text{-}t}$. The temporal embeddings of the j-th transformer E_j^t is defined as $[m_j^{agg\text{-}t}, m_j^{tmp}]$.

The input of the j-th transformer m_j^{in} is defined as the addition of E_j^f and E_j^t, and its output equals to $\Psi_j^M(m_j^{in})$. We combine the [CLS] output of all the n modality-specific transformers as the video representation V^{rep} for hierarchical attention learning and holistic ranking learning:

$$V^{rep} = \left\{ m_j^{rep} | m_j^{rep} = \Psi_j^M\left(m_j^{in}\right)^{[c]} \right\}_{j=1}^{n} \tag{2}$$

3.2 Local Alignment Module

Inspired by the Stacked Cross Attention proposed by [Lee *et al.*, 2018], we attend word embeddings T^w and video representation V^{rep} for cross attention calculation. Multi-step alignment is then conducted for hierarchical learning.

For each sample in T^w (i.e., w_i) and V^{rep} (i.e., m_j^{rep}), we first calculate the cosine similarity s_{ij} between them:

$$s_{ij} = \frac{(w_i)^T \cdot m_j^{rep}}{\|w_i\| \cdot \left\|m_j^{rep}\right\|}, \quad i \in [1, l], j \in [1, n] \tag{3}$$

Then we normalize it by using a relu function (i.e., $relu(x) = max(0, x)$) as in [Gabeur *et al.*, 2020]:

$$\overline{s}_{ij} = \frac{relu(s_{ij})}{\sqrt{\sum_{i=1}^{l} relu(s_{ij})^2}} \tag{4}$$

Attention is performed over V^{rep} given w_i in T^w, and the attended video representation vector a_i is defined as:

$$a_i = \sum_{j=1}^{n} \alpha_{ij} \cdot m_j^{rep}, \quad \alpha_{ij} = \frac{exp(\lambda \overline{s}_{ij})}{\sum_{j=1}^{n} exp(\lambda \overline{s}_{ij})} \tag{5}$$

where λ is the inverse temperature parameter of the softmax function [Chorowski *et al.*, 2015] to adjust the smoothness of the attention distribution.

Each element in $A = \{a_i\}_{i=1}^{l}$ captures related semantics shared by each word embedding w_i in T^w and the whole V^{rep}. And such context information will help to determine the shared semantics with respect to V^{rep} in the next matching step, forming a multi-step computation process.

For multi-step hierarchical alignment of knowledge in T^w and V^{rep}, an aggregating function $f(\cdot)$ is utilized to update w_i in T^w for the next alignment by aggregating them with the corresponding alignment features in A (i.e., a_i) dynamically: $w_i^* = f(w_i, a_i)$. Inspired by [Chen *et al.*, 2020b], we define our aggregating function as a modified gating mechanism:

$$\begin{aligned} f(w_i, a_i) &= g_i \cdot w_i + (1 - g_i) \cdot o_i \\ g_i &= gate(F_g[w_i, a_i] + b_g) \\ o_i &= tanh(F_o[w_i, a_i] + b_o) \end{aligned} \tag{6}$$

where g_i performs as a gate to select the most salient information, and o_i is a fused feature that enhances the interaction between w_i and a_i. F_g, F_o, b_g, and b_o are to-be-learned parameters. As defined in Equation 6, both w_i itself and o_i are utilized by the gating mechanism to refine w_i. The gate g_i can not only help to filter inconsequential information in T^w, but enable the representation learning of each w_i to focus more on its shared semantics with V^{rep}.

We combine all the equations in the above of this subsection (i.e., Equation 3 to 6) as F_a, which takes T^w and V^{rep} as inputs and is performed K times iteratively:

$$A_k, T_k^w = F_a(T_{k-1}^w, V^{rep}), \quad k \in [1, K] \tag{7}$$

where k is utilized to denote the k-th step of alignment, T_k^w and T_{k-1}^w indicate the step-wise features of word embeddings. And when $k = 0$, $T_0^w = T^w$.

In order to determine the importance of each word given the video modalities, relevance between the i-th word and the video modalities is defined as cosine similarity between the attended vector a_i and each word embedding w_i:

$$S(T^w, V^{rep}) = \sum_{k=1}^{K} \sum_{i=1}^{l} \frac{(w_{ki})^T \cdot a_{ki}}{\|w_{ki}\| \cdot \|a_{ki}\|} \tag{8}$$

The Triplet loss is utilized to enforce paired video-text pairs to be close and unpaired ones to be separated in the embedding spaces as in many previous works [Lee *et al.*, 2018; Chen *et al.*, 2020b]. Following [Faghri *et al.*, 2017], we only concentrate on the hardest negatives in a mini-batch instead of comparing with all negatives:

$$\begin{aligned} L_{Tri} = &\sum_{b=1}^{B} relu\left(\Delta - S(T_b^w, V_b^{rep}) + S(T_b^w, V_{b^*}^{rep})\right) \\ &+ \sum_{b=1}^{B} relu\left(\Delta - S(T_b^w, V_b^{rep}) + S(T_{b^*}^w, V_b^{rep})\right) \end{aligned} \tag{9}$$

where $S(T^w, V^{rep})$ is the semantic similarity between T^w and V^{rep}, B is the batch size during training. Δ is a margin value. T^w and V^{rep} with the same subscript b are paired samples. Hard negatives are indicated by the subscript b^*.

The advantage of our multi-step hierarchical alignment is that the loss function can directly supervise the learning of video-text correspondences at each matching step, helping the model to yield a higher quality of the alignment.

3.3 Holistic Ranking Module

For holistic feature learning, global information of texts and videos should be extracted and aligned.

We denote the [CLS] output of the Bert model as T^c and learn gated embedding module F_g following [Miech *et al.*, 2018] to match the size of T^c with that of videos:

$$T^g = F_g(T^c) \tag{10}$$

Since the video representation V^{rep} only present features of each modality in a video, the holistic video features need to be extracted. We utilize a holistic transformer $\Psi^H(\cdot)$ for holistic video feature learning, thus the relationships of each modality can be learned at the same time. We compute the feature embeddings of V^{rep} (i.e., $V^{agg\text{-}f}$) the same way as described in Section 3.1, and we don't calculate the temporal embeddings since temporal information is not available in V^{rep}. The concatenation of v^{rep} and its corresponding feature embeddings are fed to the holistic transformer:

$$V^h = \Psi^H([V^{agg\text{-}f}, v^{rep}]) \tag{11}$$

We utilize $S^h(T^g, V^h)$ to denote the similarity of holistic information in videos and texts, which is supervised by the bi-directional max-margin ranking loss [Karpathy *et al.*, 2014]:

$$\begin{aligned} L_{Mar} = &\sum_{b=1}^{B} \sum_{d \neq b} (relu(S^h(T_b^g, V_d^h) - S^h(T_b^g, V_b^h) + \Theta) \\ &+ relu(S^h(T_d^g, V_b^h) - S^h(T_b^g, V_b^h) + \Theta)) \end{aligned} \tag{12}$$

where Θ is the margin. Different from the Triplet loss defined in Equation 9, this loss utilizes all the samples in one mini-batch for back-propagation, and it enforces the similarity for true video-text pairs (i.e., $S^h(T_b^g, V_b^h)$) to be higher than the similarity of negative samples ($S^h(T_b^g, V_d^h)$ or $S^h(T_d^g, V_b^h)$), for all $b \neq d$, by at least Θ.

The advantage of our modality-specific-to-holistic transformer structure over the state-of-the-art MMT [Gabeur *et al.*, 2020] and Support Set [Patrick *et al.*, 2020] is that our model can learn modality-specific features (containing temporal features in each modality) and holistic features (containing relationships among all modalities) independently, rather than learning the overall features simultaneously. The learning ability of Transformers is better explored.

3.4 Overall Objective Function

The overall objective function of our model is defined as the weighted addition of the Triplet Loss L_{Tri} and bi-directional max-margin ranking loss L_{Mar}:

$$L = L_{Mar} + \beta * L_{Tri} \tag{13}$$

where β balances the two objectives.

4 Experiments

We evaluate the performance of our model on several public datasets. In the following of this section, we first introduce the datasets and metrics for result comparison, and implementation details, then illustrate the overall performance of our model on the datasets, and finally present ablation studies to evaluate the effectiveness of each component of our model.

4.1 Datasets and Metrics

HowTo100M

HowTo100M [Miech *et al.*, 2019] consists of more than one million instructional videos from YouTube. The automatically-extracted speech transcriptions are utilized to form the text for retrieval. Since this text is naturally noisy and often do not describe the visual content accurately, we only utilize this dataset for pre-training as [Gabeur *et al.*, 2020; Patrick *et al.*, 2020].

MSR-VTT

MSR-VTT [Xu *et al.*, 2016] contains ten thousand YouTube videos, each of which is paired with 20 natural sentences describing it. We follow [Yu *et al.*, 2018] and utilize the '1k-A' split, where 9,000 samples are utilized for training and the other 1,000 samples are utilized for test.

Methods	Text ⟶ Video					Video ⟶ Text				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
[Miech *et al.*, 2019]	12.1	35.0	48.0	12.0	-	-	-	-	-	-
[Liu *et al.*, 2019]	20.9	48.8	62.4	6.0	28.2	20.6	50.3	64.0	5.3	25.1
[Gabeur *et al.*, 2020]	24.6	54.0	67.1	4.0	26.7	24.4	56.0	67.8	4.0	23.6
[Patrick *et al.*, 2020]	27.4	56.3	67.7	**3.0**	-	26.6	55.1	67.5	**3.0**	-
Ours	**28.8**	**60.0**	**73.3**	**3.0**	**22.1**	**28.4**	**61.0**	**72.9**	**3.0**	**19.3**
[Zhu and Yang, 2020]*	8.6	23.4	33.1	36.0	-	-	-	-	-	-
[Miech *et al.*, 2019]*	14.9	40.2	52.8	9.0	-	16.8	41.7	55.1	8.0	-
[Luo *et al.*, 2020]*	18.7	44.4	58.9	7.0	-	-	-	-	-	-
[Gabeur *et al.*, 2020]*	26.6	57.1	69.6	4.0	24.0	27.0	57.5	69.7	3.7	21.3
[Patrick *et al.*, 2020]*	30.1	58.5	69.3	**3.0**	-	28.5	58.6	71.6	**3.0**	-
Ours*	**31.2**	**62.8**	**76.4**	**3.0**	**18.9**	**29.4**	**63.4**	**75.0**	**3.0**	**17.7**

Table 1: Retrieval performance on the MSR-VTT dataset. Methods marked with * denote they are pre-trained on the HowTo100M dataset.

Methods	Text ⟶ Video					Video ⟶ Text				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
[Liu *et al.*, 2019]	11.2	26.9	34.8	25.3	-	-	-	-	-	-
[Gabeur *et al.*, 2020]	13.2	29.2	38.8	21.0	76.3	12.1	29.3	37.9	22.5	77.1
[Gabeur *et al.*, 2020]*	12.9	29.9	40.1	19.3	75.0	12.3	28.6	38.9	20.0	76.0
Ours	**15.6**	**32.6**	**41.8**	**16.0**	**71.8**	**13.7**	**33.1**	**42.0**	**17.0**	**70.2**
Ours*	**15.8**	**34.1**	**43.6**	**14.3**	**71.4**	**14.3**	**33.7**	**43.6**	**15.5**	**68.1**

Table 2: Retrieval performance on the LSMDC dataset. Methods marked with * denote they are pre-trained on the HowTo100M dataset.

ActivityNet Captions
ActivityNet Captions [Caba Heilbron *et al.*, 2015] is composed of twenty thousand YouTube videos temporally annotated with sentence descriptions. Following the approach of [Zhang *et al.*, 2018], we concatenate all the descriptions of one video to form a paragraph. We evaluate our model on the 'val1' split, where 10,009 videos are utilized for training and 4,917 videos are utilized for test.

LSMDC
LSMDC [Torabi *et al.*, 2016] consists of 118,081 video clips extracted from 202 movies. The corresponding text of each video clip is extracted from the script or video description. 1,000 videos are utilized for test and the other videos are utilized for training.

Metrics
The performance of our model is evaluated with standard retrieval metrics used by many retrieval methods like [Lee *et al.*, 2018; Gabeur *et al.*, 2020; Patrick *et al.*, 2020]: recall at rank N (R@N, higher is better), mean rank (MnR, lower is better), and median rank (MdR, lower is better). For the MSR-VTT and LSMDC datasets, we report $N = \{1, 5, 10\}$, and for the ActivityNet dataset, we report $N = \{1, 5, 50\}$ following [Gabeur *et al.*, 2020; Patrick *et al.*, 2020]. For all the results presented by our model, we report the mean experiments with 5 random seeds.

4.2 Implementation Details

We follow [Gabeur *et al.*, 2020] and utilize seven modalities for feature encoding, which are Motion using S3D [Xie *et al.*, 2018] pre-trained on the Kinetics dataset, Audio using VGGish [Hershey *et al.*, 2017] pre-trained on the YT8M, Scene using DenseNet-161 [Huang *et al.*, 2017] pre-trained on the Places365, Appearance using SENet-154 [Hu *et al.*, 2018] pre-trained on ImageNet, and OCR, Face, and Speech using models presented by [Gabeur *et al.*, 2020]. For MSR-VTT and LSMDC datasets, we utilize all the above modalities and set $n = 7$, and for datasets that contain longer videos, e.g., HowTo100M and ActivityNet, we only utilize Motion and Audio modalities and set $n = 2$.

The max text length l of the MSR-VTT and LSMDC datasets equals to 30 and that of the ActivityNet and HowTo100M equals to 100. The inverse temperature parameter of the softmax function λ in Equation 5 equals to 9, the iteration time K in Equation 7 equals to 3, the mini-batch size B equals to 32 for the MSR-VTT, LSMDC, and ActivityNet and equals to 64 for the HowTo100M. The margin value Δ in L_{Tri} equals to 0.2, and the margin value Θ in L_{Mar} equals to 0.05. β utilized to balance the two losses equals to 1e-4 for the MSR-VTT, LSMDC, and equals to 1e-5 for the HowTo100M and ActivityNet.

The HowTo100M is only used for pre-training, and we present the results on the other three datasets. We utilize Adam optimizer for all our experiments. The modalities are used only for feature extraction rather than model training. The learning rate of all our experiments is initialized as 5e-5. We train our model on the HowTo100M for 2 million optimization steps and decay the learning rate by a multiplicative factor of 0.98 every ten thousand steps. For the MSR-VTT

Methods	Text ⟶ Video					Video ⟶ Text				
	R@1↑	R@5↑	R@50↑	MdR↓	MnR↓	R@1↑	R@5↑	R@50↑	MdR↓	MnR↓
[Liu *et al.*, 2019]	18.2	47.7	91.4	6.0	23.1	17.1	46.6	90.9	6.0	24.4
[Gabeur *et al.*, 2020]	22.7	54.2	93.2	5.0	20.8	22.9	54.8	93.1	4.3	21.2
[Patrick *et al.*, 2020]	26.8	58.1	93.5	**3.0**	-	25.5	57.3	93.5	**3.0**	-
Ours	**27.5**	**59.0**	**93.7**	**3.0**	**19.8**	**25.8**	**58.3**	**93.6**	**3.0**	**20.5**
[Liu *et al.*, 2019]*	27.3	61.1	94.4	-	-	27.9	61.6	95.0	-	-
[Wei *et al.*, 2020]*	28.5	62.6	94.9	-	-	27.9	61.9	94.1	-	-
[Gabeur *et al.*, 2020]*	28.7	61.4	94.5	3.3	16.0	28.9	61.6	94.3	4.0	17.1
[Patrick *et al.*, 2020]*	29.2	61.6	94.7	**3.0**	-	28.7	60.8	94.8	**2.0**	-
Ours*	**30.2**	**63.5**	**95.3**	**3.0**	**15.5**	**30.1**	**63.8**	**95.2**	3.0	**15.7**

Table 3: Retrieval performance on the ActivityNet dataset. Methods marked with * denote they are pre-trained on the HowTo100M dataset.

and LSMDC, we train our model for 100 thousand steps and decay the learning rate by 0.95 every one thousand steps. And for the ActivityNet, we train our model for 100 thousand steps and decay the learning rate by 0.90 every 1,000 steps.

For the text encoder in our model, we utilize the 'bert-base-uncased' checkpoint of the Bert model and fine-tune it with a dropout rate of 0.1 during training. Our modality-specific transformers and the holistic transformer are composed of 2 layers and 4 attention heads, a dropout rate of 0.1, a hidden size d_{uni} of 1024, and an intermediate size of 3072.

Our experiments are conducted on NVIDIA V100 32G GPUs. Taking the MSR-VTT dataset for example, training our model from scratch on this dataset on a single GPU takes about 15 GPU hours for 100 epochs. The average GPU hours/epochs speed of our model is about half of the speed of [Gabeur *et al.*, 2020] and quadruple of that of [Patrick *et al.*, 2020]. The inference speed of text and video bi-directional retrieval for 1000 text-video pairs is around 5 seconds.

Figure 3: Some visualization results of the weights of each modality in the holistic ranking module. As can be seen, our model can utilize more information in modalities besides "Motion".

4.3 Experimental Results

Overall Performance

We compare the bi-directional retrieval (i.e., both text-to-video and video-to-text) performance of our method to other recent works. Table 1 to 3 show results on the MSR-VTT, LSMDC, and ActivityNet datasets with and without pre-training on the HowTo100M. Without pre-training, our model outperforms all the other methods on all datasets. With pre-training, except for the median rank of video-to-text on the ActivityNet, our model outperforms other methods.

Specifically, as shown in Table 1, the results of our model on the MSR-VTT dataset without pre-training almost outperform results of [Patrick *et al.*, 2020] with pre-training. As shown in Table 2, on the LSMDC dataset, the results of our model without pre-training have already outperformed results of [Gabeur *et al.*, 2020] with pre-training, not to speak of their results without pre-training. These suggest the stronger retrieval ability of our model.

The improvement of results brings by our model on the ActivityNet is not as large as that on the MSR-VTT and LSMDC. This is because we utilize more words in each text and fewer video modalities for feature learning in the ActivityNet for fair comparisons with recent works [Liu *et al.*, 2019; Gabeur *et al.*, 2020; Patrick *et al.*, 2020], the hierarchical information alignment is harder in this case.

More Visualizations

Figure 3 and 4 are provided to show the advantages of the two modules and to prove that our method indeed tackles the issue mentioned in our motivation discussion. As in Figure3, the discrimination of certain video modality is dug out by our method for better alignment. Figure 4 shows that our local alignment module gradually focuses on important words.

Ablation Studies

In Table 4, we illustrate the results of ablation studies carried out for effectiveness evaluation of each part of our model.

To evaluate the effectiveness of modality-specific transformers (denoted as MSTs in Table 4), we replace them with the multi-modal transformer proposed by [Gabeur *et al.*, 2020]. Experimental results demonstrate that our MSTs can achieve better results, suggesting better information learning ability of our model. When the Local Alignment Module or Holistic Ranking Module (LAM or HRM in Table 4) is absent in our model, the retrieval results drop to lower values. Note that even without one module, our model can still outperform the method in [Gabeur *et al.*, 2020] as shown in Table 4. We also conduct experiments with $K = 1\&2$. As can be ob-

Methods	Text ⟶ Video					Video ⟶ Text				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
[Gabeur *et al.*, 2020]	24.6	54.0	67.1	4.0	26.7	24.4	56.0	67.8	4.0	23.6
Ours with XLNet	26.1	56.8	70.2	4.0	23.9	27.1	56.5	70.1	4.0	20.9
Ours with $n = 6$	27.2	58.5	71.6	**3.0**	23.4	27.6	59.3	70.5	**3.0**	20.5
Ours w/o MSTs	27.3	58.5	71.2	4.0	22.9	27.1	59.3	70.8	4.0	20.0
Ours w/o HRM	26.6	58.6	69.8	4.0	25.1	26.7	58.2	69.0	4.0	22.1
Ours w/o LAM	26.5	57.8	70.8	4.0	23.9	26.4	58.5	69.7	4.0	20.9
Ours with $K = 1$	27.4	59.6	71.7	3.5	22.5	27.9	60.1	71.9	3.0	20.2
Ours with $K = 2$	28.0	59.7	72.6	3.0	22.3	28.1	60.6	72.3	3.0	19.8
Ours	**28.8**	**60.0**	**73.3**	**3.0**	**22.1**	**28.4**	**61.0**	**72.9**	**3.0**	**19.3**

Table 4: Ablation studies on the MSR-VTT dataset, where MSTs denote the modality-specific transformers, LAM denotes the Local Alignment Module, HRM denotes the Holistic Ranking Module.

Figure 4: Visualizations of our local alignment module. The top left corner of the left frames show the similarities between the video and text. The top video is the ground-truth. For phrases, we report the averaged weights of each word in the phrase. As shown, the module can gradually find and focus on the important words it supposes.

served, experimental results in the last four lines in Table 4 are gradually better from top to bottom. This suggests that 1) aligning word embeddings with modality-specific video representation is better than not conducting local alignment, and 2) aligning modality-specific video representation to word embeddings with more steps brings better results.

We also analyze the limitation of our work. As [Liu *et al.*, 2019; Gabeur *et al.*, 2020; Patrick *et al.*, 2020], we utilize multiple models pre-trained on video-related large-scale datasets as experts for video feature extraction. On one hand, the ability of retrieval models is restricted by the capacity of the extracted features. On the other hand, the feature extraction and learning pipeline is rather complicated and may not be suitable to be applied to end-to-end applications. In the future, we will focus on end-to-end and more lightweight solutions for text and video bi-directional retrieval.

Which Modality Is More Important?

As described above, we utilize $n = 7$ modalities for the MSR-VTT [Xu *et al.*, 2016] and LSMDC [Torabi *et al.*, 2016] datasets, and $n = 2$ modalities for the ActivityNet [Caba Heilbron *et al.*, 2015]. It is natural to think that whether the utilized modalities are of equal importance, and if not, which modality is more important for retrieval.

Modality	Importance
Motion	0.4060
Appearance	0.2681
Audio	0.1529
Scene	0.0830
Speech	0.0516
OCR	0.0278
Face	0.0106

Table 5: The averaged importance of each modalities in the '1k-A' test set of the MSR-VTT.

After model training, we acquire the importance of the seven modalities belonging to each sample in the '1k-A' test set of the MSR-VTT, which contains 1000 videos. The averaged importance of each modality is calculated and shown in Table 5. As can be observed, Motion and Appearance are much more important than other modalities, suggesting that the two modalities are commonly described in texts and are more semantically cognizable.

More Modalities, Better Performance?

As shown in Table 5, the importance of OCR and Face is much lower than the Motion and Appearance. It is natural to think that if only utilizing the relatively more important modalities, whether the performance of our model can still be acceptable.

To evaluate this thought, we conduct an experiment with the top 6 modalities in Table 5 and without the Face modality. We utilize the same experimental settings as in Section 4.2 for setting $n = 6$ and the results are shown in Table 4 Line 3. Compared to the best results presented in the last line of the table, even without the 'least' important modality for training, the performance of the model drop by around 1-2%. This suggests that more modalities may bring better results.

Better Text Encoders, Better Performance?

Better text encoders usually brings better results. [Gabeur *et al.*, 2020] has demonstrated that Bert [Devlin *et al.*, 2018] is

better than GrOVLE [Burns *et al.*, 2019] for text encoding and [Patrick *et al.*, 2020] has demonstrated that T5 [Raffel *et al.*, 2019] is better than W2V [Pennington *et al.*, 2014] for text encoding.

However, we argue that this may not always happen. We conduct experiments with XLNet [Yang *et al.*, 2019] as the text encoder, which has been demonstrated to be more powerful than Bert on some NLP tasks. We utilize the same experimental settings as described in our manuscript except for the text encoder and the results are shown in Table 4 Line 2. Compared to the best results presented in our manuscript the model presents worse results when XLNet is utilized.

5 Conclusion

In this paper, we propose a model that first captures local semantics by utilizing multiple modality-specific transformers, and then explores hierarchical structural relationships among local semantics by multi-step video modality-word alignment, and meanwhile uses a holistic transformer to ensure the global alignment. Experimental results on three public video retrieval datasets demonstrate the advantages of our model. Ablation studies are also carried out to verify the effectiveness of each part of our model.

References

[Burns *et al.*, 2019] Andrea Burns, Reuben Tan, Kate Saenko, Stan Sclaroff, and Bryan A Plummer. Language features matter: Effective language representations for vision-language tasks. In *ICCV*, pages 7474–7483, 2019.

[Caba Heilbron *et al.*, 2015] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[Chen *et al.*, 2020a] Feiyu Chen, Jie Shao, Yonghui Zhang, et al. Interclass-relativity-adaptive metric learning for cross-modal matching and beyond. *IEEE TMM*, 2020.

[Chen *et al.*, 2020b] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *CVPR*, pages 12655–12663, 2020.

[Chen *et al.*, 2020c] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, pages 10638–10647, 2020.

[Chorowski *et al.*, 2015] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, et al. Attention-based models for speech recognition. In *NeurIPS*, pages 577–585, 2015.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2018.

[Faghri *et al.*, 2017] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, et al. VSE++: Improving visual-semantic embeddings with hard negatives. *arXiv*, 2017.

[Feng *et al.*, 2020] Zerun Feng, Zhimin Zeng, Caili Guo, and Zheng Li. Exploiting visual semantic reasoning for video-text retrieval. *arXiv*, 2020.

[Gabeur *et al.*, 2020] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020.

[Hershey *et al.*, 2017] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, pages 131–135, 2017.

[Hu *et al.*, 2018] Jie Hu, Li Shen, et al. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, et al. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

[Karpathy *et al.*, 2014] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NeurIPS*, pages 1889–1897, 2014.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.

[Li *et al.*, 2020] Zheng Li, Caili Guo, Bo Yang, Zerun Feng, and Hao Zhang. A novel convolutional architecture for video-text retrieval. In *ICME*, pages 1–6, 2020.

[Liu *et al.*, 2019] Yang Liu, Samuel Albanie, Arsha Nagrani, et al. Use what you have: Video retrieval using representations from collaborative experts. *arXiv*, 2019.

[Luo *et al.*, 2020] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, et al. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *arXiv*, 2020.

[Miech *et al.*, 2018] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv*, 2018.

[Miech *et al.*, 2019] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, et al. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019.

[Miech *et al.*, 2020] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, et al. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020.

[Patrick *et al.*, 2020] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, et al. Support-set bottlenecks for video-text representation learning. *arXiv*, 2020.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Raffel *et al.*, 2019] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv*, 2019.

[Rouditchenko *et al.*, 2020] Andrew Rouditchenko, Angie Boggust, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv*, 2020.

[Torabi *et al.*, 2016] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv*, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2020] Wei Wang, Junyu Gao, Xiaoshan Yang, and Changsheng Xu. Learning coarse-to-fine graph neural networks for video-text retrieval. *IEEE TMM*, 2020.

[Wei *et al.*, 2020] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, et al. Universal weighting metric learning for cross-moda matching. In *CVPR*, pages 13005–13014, 2020.

[Xie *et al.*, 2018] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018.

[Xu *et al.*, 2016] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5753–5763, 2019.

[Yang *et al.*, 2020] Xun Yang, Jianfeng Dong, Yixin Cao, et al. Tree-augmented cross-modal encoding for complex-query video retrieval. In *SIGIR*, pages 1339–1348, 2020.

[Yu *et al.*, 2018] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, pages 471–487, 2018.

[Zhang *et al.*, 2018] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, pages 374–390, 2018.

[Zhao *et al.*, 2020] Rui Zhao, Kecheng Zheng, and Zhengjun Zha. Stacked convolutional deep encoding network for video-text retrieval. In *ICME*, pages 1–6, 2020.

[Zhu and Yang, 2020] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755, 2020.