Computational Visual Media
https://doi.org/10.1007/s41095-0xx-xxxx-x

Research/Review Article

M2HF: Multi-branch Multi-modal Hybrid Fusion for Text-Video Retrieval

Shuo Liu[1,2*], Weize Quan[1,2], Ming Zhou[3], Sihong Chen[4](✉), Jian Kang[4], Zhe Zhao[4], Kimmo Yan[4], Chen Chen[4], and Dong-Ming Yan[1,2](✉)

Abstract Videos contain multi-modal content, and exploring multi-branch cross-modal interactions with natural language queries can provide great prominence to text-video retrieval task (TVR). However, new trending methods applying large-scale pre-trained model CLIP for TVR only focus on visual cues in videos. Furthermore, the traditional methods of simply concatenating multi-modal features do not exploit fine-grained cross-modal information in videos. In this paper, we propose a multi-branch multi-modal hybrid fusion (M2HF) network to hierarchically explore comprehensive interactions between text queries and each modality content in videos. Specifically, M2HF first utilizes visual features extracted by CLIP to early fuse with audio and motion features extracted from videos, obtaining audio-visual fusion features and motion-visual fusion features respectively. Multi-modal completion problem is also considered and solved in this process. Then, visual features, audio-visual fusion features, motion-visual fusion features, and texts extracted from videos establish cross-modal relationships with caption text queries in a multi-branch way. Finally, the retrieval outputs from all branches are late fused to obtain final text-video retrieval results. Our framework provides two kinds of training strategies, including an ensemble manner and an end-to-end manner. Moreover, a novel multi-modal balance loss function is proposed to balance the contributions of each modality for efficient end-to-end training. M2HF allows us to obtain state-of-the-art results on various benchmarks, *e.g.*, Rank@1 of 66.0%, 68.6%, 33.9%, 57.4%, 57.3% on MSR-VTT, MSVD, LSMDC, DiDeMo, and ActivityNet, respectively.

Keywords Multi-modality; Multi-branch; Hybrid Fusion; Text-Video Retrieval.

1 Introduction

With billions of videos uploaded at any time on online video platforms, it is worthwhile to retrieve the best corresponding video for a given query to efficiently access the desired video [43, 47, 48, 61]. Therefore, the tasks of Text-to-Video (T2V) and Video-to-Text (V2T) are tackled in this paper. T2V aims to obtain the ranking of all candidate videos for each caption query, while V2T finds the ranking of all candidate captions for each video query.

Unlike images, video is a kind of media owning multiple different modalities. Therefore, considering and exploring different modalities in videos is necessary for text-video retrieval. Some traditional methods [15, 18] have paid attention to this point. For example, MMT [18] first extracted different kinds of features, including audio, visual, motion, and face, etc, to obtain a richer video representation. However, they simply *concatenate* all these features and then feed them into a transformer encoder. This blind multi-modal fusion approach may cause the model to focus on certain modalities and overwhelm other informative modalities, hindering the final retrieval performance.

Recently, several methods [7, 17, 32, 34] have tried to utilize the pre-trained text-image retrieval model CLIP (contrastive language-image pretraining) [39], which is trained on 400 million text-image pairs to learn representation between text and image, as the backbone to conduct text-video retrieval task. For instance, CLIP4Clip [32] first utilized CLIP to extract the visual frame features and the caption text token features and then accumulated the similarity scores between frame-level video features and sentence-level text features for the final results. Based on CLIP4Clip, a recent parallel work Hun Yuan_tvr [34] formulated video-sentence, clip-phrase, and frame-word relationships to explore cross-modal interactions. Unfortunately, these CLIP-based works entirely ignore other rich multi-modal signals, such as audio, motion, and speech in videos.

To solve the above drawbacks, in this paper, we propose a novel method, *i.e.*, Multi-branch Multi-Modal Hybrid Fusion (M2HF), for text-video retrieval. As shown in Fig. 1, existing video retrieval works that simply concatenate the features of all modalities will ignore some weak multi-modal features but contain essential contents, such as audio, motion, and text information, since the visual features are with more semantic and extracted from stronger backbones. Therefore, our M2HF not only exploits the multiple modality information in an *explicit* multi-branch manner but also embraces the powerful pre-trained model CLIP from the perspective of multi-modal fusion. It can simultaneously exploit cross-modal relationships and tolerate the intervention and asynchrony of different modalities. First, M2HF early fuses audio and motion features respectively with visual features extracted by CLIP, producing audio-guided visual features and motion-guided visual features, which explicitly pay attention to sound sources and moving objects. Then, we comprehensively exploit the relationships between caption text queries and visual features, audio-guided visual features, motion-guided visual features, and speech contents from ASR (automatic speech recognition) in a multi-branch way. Finally, the results at each branch are combined via a fusion approach as the final retrieval result.

Our contributions can be summarized as follows:

- We propose a *Multi-branch Multi-modal Hybrid Fusion* network to solve text-video retrieval tasks, where our method achieves state-of-the-art Rank@1 retrieval results on five public benchmarks.
- Instead of simply concatenating all modality features, *M2HF* exploits the multiple modality contents in an explicit multi-branch manner when some of these features are weak and some are powerful, and organically integrates these modalities to improve the performance of text-video retrieval task remarkably.
- *M2HF* is trained by two strategies: end-to-end training (E2E) and an ensemble manner (Ensemble), where a novel *multi-modal balance loss* is designed for E2E. This loss can balance the branches and optimize the entire network training.

2 Related Work

2.1 Multi-modal Fusion

2.1.1 Early Fusion

Such methods mainly fuse multiple modalities at the feature branch. Bilinear pooling-based approaches fuse two modalities by learning a joint representation space, *e.g.*, MLB (low-rank bilinear pooling) [26] and MFB (multi-modal factorized bilinear pooling) [59], etc. Attention-based methods fuse features from different modalities based on the correlation, including channel-wise attention [25], non-local model [49], and transformer [44, 56], etc. Some researches [45, 62, 63] apply early fusion methods to solve multi-modal visual media editing and tracking tasks. The Write-A-Video [45] is an innovative tool designed for creating video montages primarily through text editing. By inputting themed text and selecting a related video repository, this tool enables users to effortlessly generate video montages by automatically searching semantically matching candidate shots and optimization-based shot assembly. Zhang *et al.* [62] introduced a 3D animation system to create, edit, preview, and render animations solely through text editing. This system initially parses these texts into semantic scene graphs and then retrieves 3D object models for the composition of virtual scenes and motion clips for character animation. Zhang *et al.* [63] delivered a comprehensive review of multi-modal tracking, culminating in the integration of multi-modal trackers within a unified framework.

2.1.2 Late Fusion

Late fusion, also known as decision-branch fusion, first trains different models on different modalities and then fuses the predictions of these trained models [42]. Late fusion methods mainly design different combination strategies to merge models' outputs, such as voting, average combination, ensemble learning, and other combination methods.

2.1.3 Hybrid Fusion

Hybrid fusion [6, 36, 51, 54] attempts to exploit the advantages of both early fusion and late fusion. Chen *et al.* [11] proposed a novel approach that integrates both early and late fusion techniques within a single model. Initially, they trained deep Bidirectional Long Short-Term Memory (Bi-LSTM) networks to process unimodal features. Subsequently, these Bi-LSTMs were employed to fuse bi-modal features at an early stage. Ultimately, the outputs from various models were amalgamated at a later stage using a second-level Bi-LSTM, enhancing the prediction accuracy. Hybrid multimodal fusion has been used for emotion recognition [33]. It uses an early fusion strategy to obtain the audio-visual features and

[1] MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: S.Liu, liushuo2019@ia.ac.cn; W.Quan, qweizework@gmail.com; D.-M. Yan, yandongming@gmail.com.

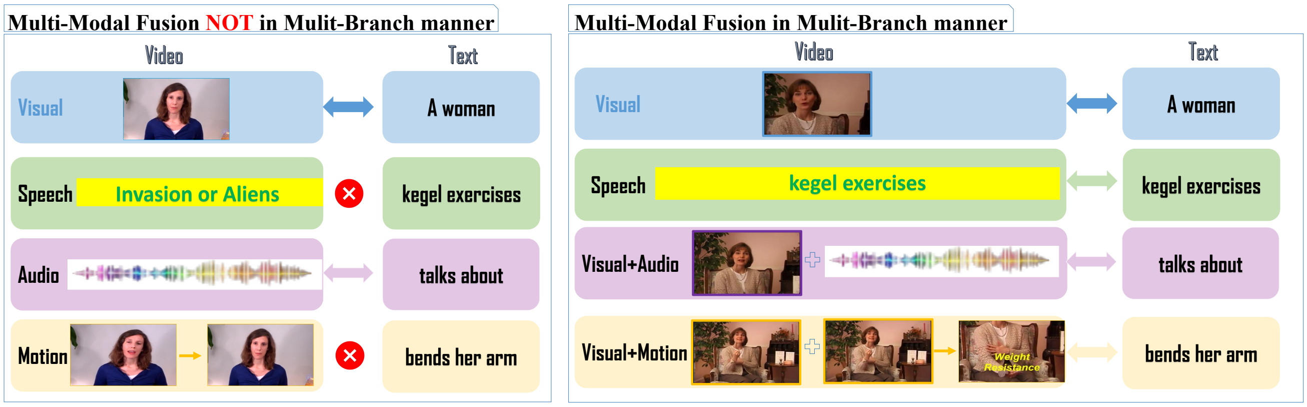
[2] the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.

[3] Donghua University, Shanghai 200051, China. E-mail: M.Zhou, mzhou@mail.dhu.edu.cn.

[4] Tencent, Shenzhen 518057, China. E-mail: S.Chen, cshwhale@sina.com; J.Kang, kangjianqh@sina.com; Z.Zhao, nlpzhezhao@tencent.com; K.Yan, kimmoyan@tencent.com; C.Chen, chen1634chen@gmail.com.

[*] Work done when Shuo Liu interned at Tencent.

Manuscript received: 2022-01-01; accepted: 2022-01-01



(a) Previous methods cannot tolerate the intervention and asynchrony of different modalities.

(b) Our M2HF can fully and effectively exploit multimodal content.

Fig. 1 The comparison between the previous multi-modal fusion methods which are not in a multi-branch manner (a) and our multi-modal fusion method in a multi-branch manner (b). Each branch ("Visual", "Audio", "Speech", and "Motion") of M2HF from videos can fully explore the multi-modal contents and build more explicit relationships with germane text tokens of captions for better video understanding and retrieval performance compared to previous methods. Moreover, when some of these multi-modal features are powerful (Visual) and some are weak (Audio, Motion, and Speech), M2HF can fully exploit the multiple modality contents compared to previous methods.

bisignal features and a late fusion strategy to further boost the model's recognition performance. Xu *et al.* [55] uses the hybrid fusion strategy for Humor Detection to improve the model's performance.

It has been also used for multi-modal speaker identification [52] and multi-media event detection (MED) [29]. Atrey *et al.* [3] adopted a Bayesian inference fusion approach at hybrid levels (the feature and the decision levels). Ayache *et al.* [4] proposed a hybrid fusion approach to normalize early fusion and contextual late fusion for semantic indexing of multimedia resources. Lan *et al.* [29] applied the hybrid fusion method to solve multi-media event detection. This work integrates the advantages of early fusion to capture feature relationships and late fusion to handle overfitting. In the fusion process of video and sound signals in [20], the listening deep model based only on video signals and only on sound signals is first trained to generate model test results respectively. Then the integrated features of video and sound signals are input into the audio-visual system. Finally, a weighted method is used to integrate the predictions of multiple models to obtain a better recognition result. The combination strategy of hybrid fusion methods is a key factor in improving model performance. Morales *et al.* [35] trained separate prediction models for each modality and then obtained predictions from every modality. These new feature vectors are used to train a new model for the final prediction. Alghowinem *et al.* [1] connected the results of each modality to early fusion feature vectors and then performed model predictions. A majority voting method is used to evaluate the final effect. Shalu *et al.* [41] inputted the three modals' features into linear layers to obtain the corresponding scores. Then, they spliced the obtained scores to obtain the fused features and finally inputted the fused features into linear layers for model training.

Overall, the hybrid fusion strategy combines feature and decision levels, which can utilize the advantages of both early and late fusion strategies. Therefore, in this work, we also adopt the hybrid fusion mechanism. We apply this method to design a novel framework for video retrieval task to better fuse these multi-modal features when some of these features are weak and some are powerful. M2HF can simultaneously exploit cross-modal relationships and tolerate the intervention and asynchrony of different modalities.

2.2 Text-Video Retrieval

For TVR, two research directions mainly exist: multi-modal features and large-scale pre-trained models.

2.2.1 Text-Video Retrieval based on Multi-modal Features

One direction applies rich multi-modal cues to retrieve videos. MMT encoded seven modalities such as audio, visual, and motion separately, and then fed them into a transformer for better video representation. MDMMT [15] extended MMT by optimizing training datasets. MDMMT-2 [28] introduced a three-stage training process and double positional encoding for better retrieval performance. However, these methods mainly input various multi-modal features into an encoder producing video representations. This fusion method is somewhat simple and the interactions between multiple modalities and text queries are blind, potentially limiting the final retrieval performance. Instead, our fine-grained hybrid fusion method can fuse multi-modal features and explicitly model text-video relationships in a multi-branch manner.

2.2.2 Text-Video Retrieval based on CLIP

Another direction attempts to utilize pre-trained CLIP as the backbone for TVR task. The seminal work CLIP4Clip exploited CLIP to extract features of visual frames and captions, and then computed the similarity scores between video and text features. Based on CLIP, Fang *et al.* [16] introduced temporal difference block and temporal completion block to enhance temporal relationships between video frames and video captions. Cheng *et al.* [12] proposed a novel dual Softmax loss (DSL). Wang *et al.* [46] carefully studied the cross-modality interaction process and representation learning for TVR, and proposed a disentangled framework, including a weighted token-wise interaction (WTI) block and a channel decorrelation regularization block, to model the sequential and hierarchical representation. Recently, Gorti *et al.* [21] leveraged CLIP as a backbone and proposed a parametric text-conditioned pooling to aggregate video frame representations based on the similarity between the video frame and text. Hu *et al.* [24] performed feature fusion at both video and text ends with different feature extraction. Lin *et al.* [30] generated multiple prototypes for video automatically to account for its rich information, and proposed a text-adaptive matching strategy to dig the correspondence between texts and videos. Bain *et al.* [5] introduced a dual encoder model for end-to-end text-video retrieval training, taking advantage of large-scale image and video captioning datasets. However, these CLIP-based methods mainly focus on the visual modality, while ignoring other information in videos, such as motion, audio, and speech, which are still important cues for TVR task.

3 Proposed Method

3.1 Overall Architecture

Fig. 2 illustrates the entire pipeline of our M2HF. Given a set of video-text pairs $\{(V_1, t_1), ..., (V_n, t_n)\}$, our method measures the similarity of video and text from four branches, as shown in the middle of Fig. 2. M2HF explicitly establishes relationships and conducts similarity computation between text t_j and visual v_i, audio a_i, motion m_i, and speech s_i extracted from video V_i, respectively.

Multi-modal fusion is in a hybrid fusion way: audio and motion features are early fused with visual features, *i.e.*, audio-visual fusion and motion-visual fusion in Fig. 2, which mitigates the representation ability gap between visual features and other modality features to further enhance multi-modal understandings; all branchs' ranking results are late fused for the final retrieval results by selecting the best ranking in the output of each branch. We aggregate multi-modal cues in a hierarchical manner for more accurate retrieval performance. Furthermore, two training strategies (E2E and Ensemble) are provided in this paper, and a novel multi-modal balance loss is proposed to serve E2E training by minimizing each pair score and calculating the balanced loss. In the following, we describe the details of the four branches and training objectives.

3.2 Visual-to-Text Branch

The visual-to-text branch is designed for making the cross-modality relationship between visual features from video and query text features. Image encoder ViT [14] and text encoder Bert [13] of CLIP is first used to extract visual features ($v_i \in \mathbb{R}^{F \times d_v}$) and text features ($t_i \in \mathbb{R}^{T \times d_t}$), respectively, where F is the number of video frames, T is the number of text tokens, d_v and d_t represents the dimensions of visual and text features, respectively. To compute the similarity matrix $\mathcal{S}_{v-t}$ between visual features and text features, we choose the weighted token-wise interaction (WTI) function. The entire process is computed as:

$$\begin{aligned} t2v\_logits &= \sum_{p=1}^{T} f_{tw,\theta}^{p}(t_i) max_{q=1}^{F}\left(\frac{t_i^p}{\|t_i^p\|_2}\right)^{\mathrm{T}}\left(\frac{v_i^q}{\|v_i^q\|_2}\right), \\ v2t\_logits &= \sum_{q=1}^{F} f_{vw,\theta}^{q}(v_i) max_{p=1}^{T}\left(\frac{t_i^p}{\|t_i^p\|_2}\right)^{\mathrm{T}}\left(\frac{v_i^q}{\|v_i^q\|_2}\right), \\ \mathcal{S}_{v-t}[i,i] &= \mathrm{WTI}(v_i, t_i) = \frac{t2v\_logits + v2t\_logits}{2.0}, \end{aligned} \tag{1}$$

where $f_{tw,\theta}$ and $f_{vw,\theta}$ both are the combination of the MLP (multilayer perceptron) and a Softmax, i is a sampled index, p and q denote the index of text token and video frame.

3.3 Audio-to-Text Branch

At the audio-to-text branch, audio features and visual features are early fused to highlight the visual semantic information related to audio, *e.g.*, sound-producing objects. Then, the audio-guided visual features are used to build connections with text features. Audio features ($a_i \in \mathbb{R}^{F \times d_a}$) are extracted from the log mel-spectrogram via the VGGish [23] pre-trained on AudioSet [19], where d_a is the dimension of audio features. We adopt the MFB-based method in text-to-video task to early fuse audio and visual features, yielding high-level semantic audio-visual fusion features $\mathcal{F}_{av_i} \in \mathbb{R}^{F \times d_v}$. Specifically, audio features a_i and visual features v_i are projected and completed as the same dimension kd using linear layers and ReLU. The completed audio and visual features are multiplied

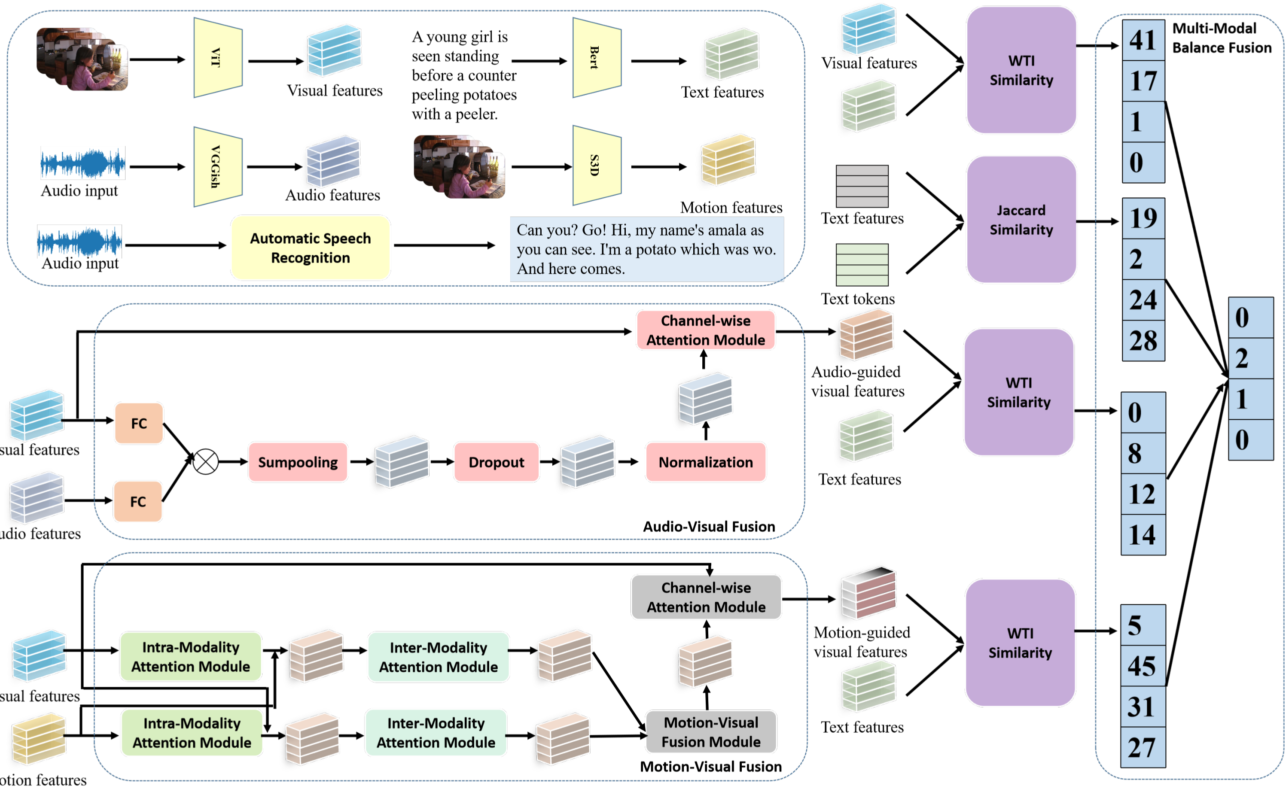

ViT
Visual features
A young girl is seen standing before a counter peeling potatoes with a peeler.
Bert
Text features
Audio input
VGGish
Audio features
S3D
Motion features
Audio input
Automatic Speech Recognition
Can you? Go! Hi, my name's amala as you can see. I'm a potato which was wo. And here comes.
Visual features
Audio features
FC
FC
Sumpooling
Dropout
Normalization
Channel-wise Attention Module
Audio-Visual Fusion
Visual features
Motion features
Intra-Modality Attention Module
Intra-Modality Attention Module
Inter-Modality Attention Module
Inter-Modality Attention Module
Channel-wise Attention Module
Motion-Visual Fusion Module
Motion-Visual Fusion
Visual features
Text features
Text tokens
Audio-guided visual features
Text features
Motion-guided visual features
Text features
WTI Similarity
Jaccard Similarity
WTI Similarity
WTI Similarity
Multi-Modal Balance Fusion
41
17
1
0
19
2
24
28
0
8
12
14
5
45
31
27
0
2
1
0

Fig. 2 The architecture of our multi-branch multi-modal hybrid fusion network (M2HF) for text-video retrieval. In the beginning, "Visual", "Audio", "Speech","Motion", and "Text" are extracted. The early fusion strategies ("Audio-Visual Fusion" and "Motion-Visual Fusion") are conducted to fuse multi-modal features with visual features. The multi-modal contents of each branch make a relation with the caption text. The late fusion strategy (Multi-modal balance fusion) is conducted to select the best ranking from all branches for the final retrieval results.

and fed into the sum pooling layer with the kernel size k. The formulation is as follows:

$$\mathcal{F}_{av_i} = \text{Drop}(\text{SP}(\Psi^{\text{T}} a_i \odot \Phi^{\text{T}} v_i, k)), \tag{2}$$

where $\Psi \in \mathbb{R}^{d_a \times (kd)}$ and $\Phi \in \mathbb{R}^{d_v \times (kd)}$ are two learnable matrices, $\odot$ represents element-wise product, $\text{SP}(\cdot, k)$ is the sum pooling with kernel size k and stride k, and $\text{Drop}(\cdot)$ is a dropout layer to prevent the over-fitting. To stabilize the model training, power and L_2 normalizations are utilized:

$$\mathcal{F}_{av_i} \leftarrow \text{sign}(\mathcal{F}_{av_i})\sqrt{|\mathcal{F}_{av_i}|}, \mathcal{F}_{av_i} \leftarrow \mathcal{F}_{av_i} / \left\|\mathcal{F}_{av_i}\right\|. \tag{3}$$

Next, the audio-visual fusion feature guides the raw visual features v_i by channel-wise attention operation for obtaining the final audio-guided visual features. A squeeze-and-excitation operation [25] is applied to produce channel-wise attentive weights ($\mathcal{W}_i^{\mathcal{A}} \in \mathbb{R}^{d_v \times 1}$). This process is formulated as:

$$\mathcal{W}_i^{\mathcal{A}} = \delta(\mathbf{W}_2 \sigma(\mathbf{W}_1(\mathcal{F}_{av_i}))), \tag{4}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_v \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d \times d_v}$ are two linear transformations with $d = \frac{d_v}{2}$; σ and δ denote the ReLU and sigmoid functions, respectively.

The final audio-guided visual features are obtained via:

$$av_i = \mathcal{W}_i^{\mathcal{A}} \odot v_i. \tag{5}$$

Similar to the visual-to-text branch, the relationship between audio-guided visual features and text features is formulated by WTI. The detailed formula of the similarity matrix $\mathcal{S}_{a-t}$ is similar with Eq.(1) replacing v_i with av_i.

In real-world scenarios, not all videos have audio signals, *i.e.*, modality missing problem, therefore, we propose a simple and effective completion method. Specifically, we pad missing audio features with element 1. The primary idea of this completion strategy is that the guidance mechanism can still work by guiding with original visual features.

3.4 Motion-to-Text Branch

The motion-to-text branch is proposed to early fuse motion features with visual features obtaining motion-guided visual features, which explicitly consider the object movement in the visual modality. The fused features then compute the similarity with text features. Similar to MMT [63], we directly use the motion features extracted from S3D [53] with the RGB sequences as input. RGB frames are also fed into ViT for extracting the visual features, however, ViT embeds each 2D frame patch independently, which ignores the temporal information among image frames. The motion features extracted with S3D can provide more spatio-temporal cues of moving objects for the M2HF framework. In our work, the motion features is defined as $m_i \in \mathbb{R}^{F \times d_m}$, where d_m is the dimension of motion features.

For the fusion of motion features and visual features, we utilize the encoder of transformer block. The detailed calculation process is as follows:

$$\begin{aligned} &\mathrm{Encoder}(Q,K,V) = \mathrm{LN}(X+Y), \\ &X = \mathrm{MHA}(\tilde{Q},\tilde{K},\tilde{V}), Y = \mathrm{FFN}(\mathrm{LN}(X+\tilde{Q})), \\ &\tilde{Q} = Q\mathbf{W}_{\tilde{Q}}, \tilde{K} = K\mathbf{W}_{\tilde{K}}, \tilde{V} = V\mathbf{W}_{\tilde{V}}, \end{aligned} \tag{6}$$

where Q, K, $V \in \mathbb{R}^{F \times d}$ are input features of transformer's encoder; $\mathbf{W}_{\tilde{Q}}, \mathbf{W}_{\tilde{K}}, \mathbf{W}_{\tilde{V}} \in \mathbb{R}^{d \times d}$ are projection matrices; LN refers to the layer normalization; MHA is the multi-head attention with 4 heads; and FFN is the feed forward network.

As shown in the bottom of Fig. 2, motion features m_i and visual features v_i are first fed into the intra-modality attention module to learn the informative segments of each modality. The motion modality is taken as an example to explain the intra-modality attention module. Specifically, motion features are first projected yielding query features ($Q \in \mathbb{R}^{F \times d_m}$), key features ($K \in \mathbb{R}^{F \times d_m}$), and value features ($V \in \mathbb{R}^{F \times d_m}$). They are then fed into the encoder of the transformer producing the self-attentive motion features $m^{self} = \mathrm{Encoder}(Q,K,V)$ via Eq. 6. Self-attentive visual features v^{self} are obtained using the same way.

Next, the inter-modality attention module is introduced to exploit the relationship between motion and visual features via the encoder of the transformer as well. Different from the intra-modality computation, key and value features of the inter-modality attention are the concatenation of one modality features and the self-attentive features of another modality. Cross-modality features m^{cross} and v^{cross} are obtained as:

$$\begin{aligned} m^{cross} &= \mathrm{Encoder}(m_i, \mathrm{cat}(m_i, v_i^{self}), \mathrm{cat}(m_i, v_i^{self})), \\ v^{cross} &= \mathrm{Encoder}(v_i, \mathrm{cat}(v_i, m_i^{self}), \mathrm{cat}(v_i, m_i^{self})), \end{aligned} \tag{7}$$

where cat is the concatenation of two features in the temporal dimension. The cross-modality features are then integrated by the motion-visual fusion module, also an encoder of the transformer, to yield motion-visual fusion features $F_{mv_i} \in \mathbb{R}^{F \times d_v}$ via:

$$\begin{aligned} &F_{mv_i} = \mathrm{Encoder}(Q,K,V), \\ &Q = m_i^{cross} \odot v_i^{cross}, K, V = \mathrm{cat}(m_i^{cross}, v_i^{cross}). \end{aligned} \tag{8}$$

After that, F_{mv_i} is used to guide the visual features to highlight the moving objects. The guidance weights are first estimated via the squeeze-and-excitation block [25] as follows:

$$\mathcal{W}_i^{\mathcal{M}} = \delta(\mathbf{W}_4 \sigma(\mathbf{W}_3(\mathcal{F}_{mv_i}))), \tag{9}$$

where $\mathbf{W}_3 \in \mathbb{R}^{d_v \times d}$ and $\mathbf{W}_4 \in \mathbb{R}^{d \times d_v}$ are two linear transformations with $d = \frac{d_v}{2}$. Motion-guided visual features mv_i are achieved via:

$$mv_i = \mathcal{W}_i^{\mathcal{M}} \odot v_i, \tag{10}$$

Finally, the similarity matrix $\mathcal{S}_{m-t}$ between motion-guided visual features and text features are calculated via the WTI, *i.e.*, replacing v_i in Eq.(1) with mv_i.

3.5 Speech-to-Text Branch

In addition to visual and motion information, videos also contain textual content, *e.g.*, subtitles and text in image frames, and text extracted from speech through ASR techniques. Considering the generality, in this work, we mainly exploit the text information related to the speech audio channel to enhance the text-video retrieval performance. Clues hidden in subtitles and text will be studied in our future work. For the speech-to-text branch, the same modality can directly compute the similarity matrix $\mathcal{S}_{s-t}$ without intervention from other modalities. Jaccard scores [38] are formulated for each pair of speech and text. Before the formulation, several pre-processing operations are conducted. First, stop words including pronouns, integrated nouns, and other less representative words are filtered from speech and text. Then, the remaining words will be filtered again to keep only nouns since nouns are more representative than verbs, adverbs, prepositions, and others. Next, the filtered tokens are converted to the same root. Finally, all the letters are lowercase yielding the final set of speech S_s and text S_t. The calculation of the Jaccard score is as:

$$\mathcal{S}_{s-t}[i,i] = \mathrm{Jaccard}(S_s, S_t) = \frac{\mathrm{len}(S_s \cap S_t)}{\mathrm{len}(S_s \cup S_t)}, \tag{11}$$

where $\mathrm{Jaccard}(\cdot,\cdot)$ computes the jaccard correlation, and $\mathrm{len}(\cdot)$ calculates the number of each set's tokens.

In fact, we have tried to extract the speech and text features and compute the similarities between them to obtain the results of this branch. However, the performance of computing the similarity with deep features is not better than that of explicitly matching the tokens from speech and text. The backbones used to extract those features depend on the attention mechanism. This mechanism drowns out keywords that appear relatively rarely, whether global or local, wasting those leads. However, the word-matching strategy cannot sacrifice these words. This branch will be developed and improved in the future.

3.6 Ensemble and E2E Text-Video Retrieval

In our work, M2HF is trained with two strategies: an ensemble manner (Ensemble) and end-to-end training (E2E). As for Ensemble, we train each branch separately obtaining the respective model and then ensemble the retrieval outputs of each model. As for E2E, we train all branches together with the help of multi-modal balance loss outputting one model, and we use the retrieval outputs of each branch to obtain final results.

3.6.1 Ensemble Retrieval

Inspired by ensemble learning, we first train the model of each multi-modality branch independently and then fuse their predictions with a late fusion method. Dual softmax loss (DSL) [12] is applied for the visual-to-text, audio-to-text, and motion-to-text branch. DSL pursues the dual optimal match and thus obtains good retrieval performance. Specifically, the similarity matrices $\mathcal{S}_{v-t}$, $\mathcal{S}_{a-t}$, and $\mathcal{S}_{m-t}$ are fed into DSL function. We take $\mathcal{S}_{v-t}$ as an example, the loss of visual-to-text branch ($\mathcal{L}_v = -\frac{1}{B}\sum_i^B \mathbf{L}_v$) formulates as follows:

$$\begin{aligned}
P_{v2t}[i,j] &= \frac{e^{(\lambda \mathcal{S}_{v-t}[i,i])}}{\sum_{j=1}^{B} e^{(\lambda \mathcal{S}_{v-t}[j,i])}}, \\
P_{t2v}[i,j] &= \frac{e^{(\lambda \mathcal{S}_{v-t}[i,i])}}{\sum_{j=1}^{B} e^{(\lambda \mathcal{S}_{v-t}[i,j])}}, \\
\mathbf{L}_{v2t}[i] &= log(\frac{e^{(\eta \mathcal{S}_{v-t}[i,i] P_{v2t}[i,i])}}{\sum_{j=1}^{B} e^{(\eta \mathcal{S}_{v-t}[i,j] P_{v2t}[i,j])}}), \\
\mathbf{L}_v &= \mathbf{L}_{v2t} + \mathbf{L}_{t2v},
\end{aligned} \tag{12}$$

where λ is a temperature hyper-parameter to smooth the gradients, B is the batch size, and η is a logit scaling factor. $\mathcal{L}_a$ and $\mathcal{L}_m$ are obtained in the same way.

For the evaluation, a novel late fusion strategy, called multi-modal balance fusion (MMBF), is proposed to aggregate the outputs of all four branches by selecting the best ranking from each branch. The ranking of each branch is denoted as $\mathcal{R}_{v-t}$, $\mathcal{R}_{a-t}$, $\mathcal{R}_{m-t}$, and $\mathcal{R}_{s-t}$, which are obtained based on the respective similarity matrices. Then, the final ranking is

$$\mathcal{R} = min(\mathcal{R}_{v-t}, \mathcal{R}_{a-t}, \mathcal{R}_{m-t}, \mathcal{R}_{s-t}), \tag{13}$$

where min is element-wise minimizing operation.

In practice, it rarely happens that different videos have the same similarity to the same retrieved text. Moreover, the hybrid fusion strategy can mitigate the situation where two videos have the same minimal rank. For example, if there is one more best choice for a case based on the $\mathcal{R}_{v-t}$ ranking, current works will choose the video with the smallest id number as the correct answer. However, our method uses four rankings to obtain the best choice, which is more robust.

3.6.2 E2E Retrieval

In addition, we introduce a novel multi-modal balance loss (MMBL) to train the model in an end-to-end manner. Specifically, MMBL minimizes each pair score in each branch yielding the final balanced loss as follows:

$$\mathcal{L} = -\frac{1}{B}\sum_{i}^{B} min(\mathbf{L}_v, \mathbf{L}_a, \mathbf{L}_m). \tag{14}$$

where B is the batch size, and min is element-wise minimizing operation. We also try other similar fusion methods, including average, element-wise maximizing, and element-wise adding, and find that the element-wise minimizing achieves the best performance. The evaluation of E2E retrieval also uses MMBF.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

In this work, we use five common benchmarks: MSR-VTT [57], MSVD [9], LSMDC [40], DiDeMo [2], and ActivityNet [27] to conduct extensive experiments to validate our method. The same settings as CLIP4Clip are used in this work.

MSR-VTT is a large-scale dataset containing 10,000 video clips and each video clip is described with 20 natural sentences via Amazon Mechanical Turks. Following the setting [58], 9,000 and 1,000 videos are used for training and testing, respectively.

MSVD has 1,970 video clips, and each video clip contains about 40 sentences. We adopt the original data split, 1,200 videos for training, 100 videos for validation, and 670 videos for testing.

LSMDC is composed of 118,081 video clips extracted from 202 movies and each video clip has a caption. The validation set and evaluation set contains 7,408 and 1,000 videos, respectively.

ActivityNet contains 20,000 YouTube videos with 100k captions. Standard split, the training set has 10,009 videos and the validation set has 4,917 videos, is followed. Like [60], we concatenate all the captions of a video as a paragraph.

DiDeMo contains 10,000 videos annotated with 40,000 sentences. All captions of a video are concatenated into a paragraph [31].

4.1.2 Metrics

For the performance evaluation, we adopt the standard retrieval metrics: Recall at rank N (R@N, higher is better), mean rank (MnR, lower is better), and median rank (MdR, lower is better).

Table 1 Retrieval results on MSR-VTT 1K dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
T2VLAD	29.5	59.0	70.1	4.0	-	31.8	60.0	71.1	3.0	-
CLIP4Clip	44.5	71.4	81.6	2.0	15.3	42.7	70.9	80.6	2.0	-
VCM	43.8	71.0	80.9	2.0	14.3	45.1	72.3	82.3	2.0	10.7
X-Pool	46.9	72.8	82.2	2.0	14.3	-	-	-	-	-
CAMOE	48.8	75.6	85.3	2.0	12.4	50.3	74.6	83.8	2.0	9.9
DCR	53.3	80.3	87.6	1.0	-	56.2	79.9	87.4	1.0	-
LAFF	45.8	71.5	82.0	-	-	-	-	-	-	-
TVMM	36.2	64.2	75.7	3.0	-	34.8	63.8	73.7	3.0	-
Hun Yuan_tvr (ViT-B/16)	55.0	80.4	86.8	1.0	10.3	55.5	78.4	85.8	1.0	7.7
Hun Yuan_tvr (ViT-L/14)	53.2	77.6	83.9	1.0	10.1	54.0	78.8	87.1	1.0	8.3
MMT (ViT-B/16)	56.3	83.4	87.1	1.0	9.2	56.1	76.1	82.4	1.0	8.2
MMT (ViT-L/14)	55.9	79.2	81.5	1.0	9.3	56.4	72.1	89.1	1.0	7.5
Ours_Ensemble (ViT-B/16)	59.9	82.8	89.3	1.0	8.1	60.7	82.7	90.2	1.0	5.6
Ours_Ensemble (ViT-L/14)	65.3	85.3	**91.7**	1.0	7.9	66.0	86.0	91.7	1.0	**5.0**
Ours_E2E (ViT-B/16)	62.1	84.8	90.7	1.0	**6.5**	63.5	85.6	91.6	1.0	5.6
Ours_E2E (ViT-L/14)	**66.0**	**86.3**	91.5	**1.0**	6.7	**65.7**	**86.2**	**91.9**	**1.0**	5.1

Table 2 Retrieval results on MSVD dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
CLIP4Clip	46.2	76.1	84.6	2.0	10.0	48.4	70.3	77.2	2.0	-
X-Pool	47.2	77.4	86.0	2.0	9.3	-	-	-	-	-
CAMOE	49.8	79.2	87.0	-	9.4	-	-	-	-	-
DCR	50.0	81.5	89.5	2.0	-	58.7	92.5	95.6	1.0	-
LAFF	45.4	75.5	84.1	-	-	-	-	-	-	-
TVMM	36.7	67.4	81.3	2.5	-	-	-	-	-	-
Hun Yuan_tvr (ViT-B/16)	54.6	82.4	89.6	1.0	8.0	58.0	85.4	89.6	1.0	5.5
Hun Yuan_tvr (ViT-L/14)	57.8	83.3	89.6	1.0	7.8	63.4	88.1	92.6	1.0	3.3
MMT (ViT-B/16)	57.1	84.1	90.4	1.0	7.5	58.5	88.1	90.0	1.0	4.2
MMT (ViT-L/14)	60.4	87.1	90.8	1.0	6.1	67.1	90.2	91.2	1.0	3.0
Ours_Ensemble (ViT-B/16)	61.7	86.1	91.5	1.0	6.3	69.1	87.3	93.9	1.0	3.4
Ours_Ensemble (ViT-L/14)	67.1	88.1	92.7	1.0	5.5	72.1	89.3	95.5	1.0	2.6
Ours_E2E (ViT-B/16)	62.7	86.3	91.7	1.0	6.0	73.3	90.6	94.3	1.0	2.7
Ours_E2E (ViT-L/14)	**68.6**	**88.7**	**92.9**	**1.0**	**5.1**	**75.7**	**91.0**	**96.3**	**1.0**	**2.6**

Table 3 Retrieval results on LSMDC dataset.

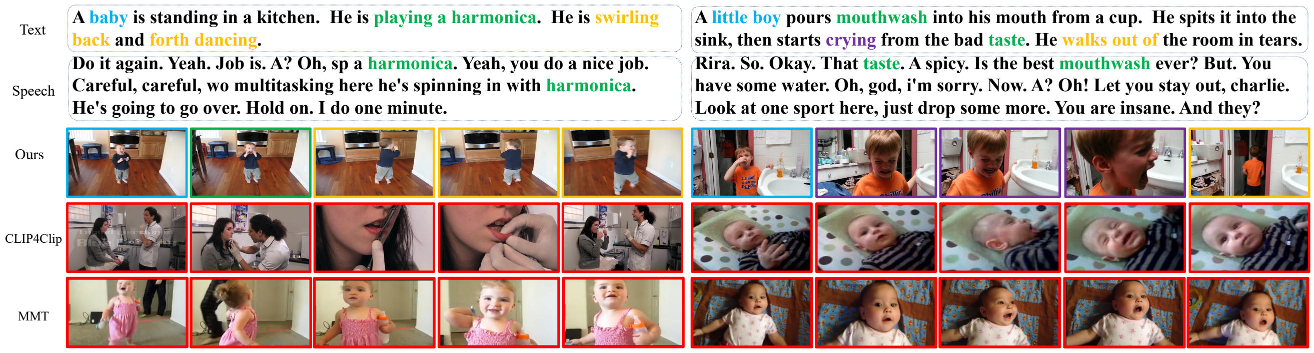
Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
T2VLAD	14.3	32.4	-	16.0	-	14.2	33.5	-	17.0	-
CLIP4Clip	22.6	41.0	49.1	11.0	-	-	-	-	-	-
X-Pool	25.2	43.7	53.5	8.0	53.2	-	-	-	-	-
CAMOE	25.9	46.1	53.7	-	54.4	-	-	-	-	-
DCR	26.5	47.6	56.8	7.0	-	27.0	45.7	55.4	8.0	-
TVMM	17.8	37.1	45.9	13.5	-	16.5	34.3	44.6	14.0	-
Hun Yuan_tvr (ViT-B/16)	26.3	46.1	54.1	7.0	55.3	27.1	46.6	54.5	7.0	45.7
Hun Yuan_tvr (ViT-L/14)	29.7	46.4	55.4	7.0	56.4	30.1	47.5	55.7	7.0	48.9
MMT (ViT-B/16)	28.1	46.3	56.5	7.0	50.3	30.1	49.3	58.1	7.0	40.4
MMT (ViT-L/14)	28.3	48.1	57.8	7.0	53.1	30.6	48.6	56.1	7.0	47.3
Ours_Ensemble (ViT-B/16)	30.2	46.8	58.2	7.0	43.2	29.8	48.1	57.8	6.0	37.2
Ours_Ensemble (ViT-L/14)	**33.9**	**55.9**	**64.2**	**4.0**	**34.8**	**34.0**	**57.1**	**64.6**	**3.0**	**28.3**
Ours_E2E (ViT-B/16)	30.3	46.4	55.9	7.0	44.8	29.1	46.8	56.3	6.0	39.6
Ours_E2E (ViT-L/14)	31.8	53.2	62.2	4.0	36.9	31.6	53.3	63.6	5.0	30.8

Table 4 Retrieval results on DiDeMo dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
CLIP4Clip	41.4	58.2	79.1	2.0	-	42.8	69.8	79.0	2.0	-
CAMOE	43.8	71.4	-	-	-	45.5	71.2	-	-	-
DCR	49.0	76.5	84.5	2.0	-	49.9	75.4	83.3	2.0	-
TVMM	36.5	64.9	75.4	3.0	-	-	-	-	-	-
Hun Yuan_tvr (ViT-B/16)	52.1	78.2	85.7	1.0	11.1	54.8	79.9	87.2	1.0	7.3
Hun Yuan_tvr (ViT-L/14)	49.5	73.7	81.6	2.0	14.8	50.3	76.5	83.7	1.0	10.4
MMT (ViT-B/16)	54.3	79.5	86.1	1.0	10.5	55.5	80.3	88.4	1.0	7.3
MMT (ViT-L/14)	53.2	72.1	84.9	2.0	16.4	54.6	78.3	85.2	1.0	9.2
Ours_Ensemble (ViT-B/16)	54.8	79.5	85.3	1.0	10.0	56.0	79.0	86.1	1.0	7.3
Ours_Ensemble (ViT-L/14)	57.1	79.6	87.3	1.0	9.6	58.4	80.6	88.8	1.0	7.3
Ours_E2E (ViT-B/16)	56.9	79.3	85.3	1.0	9.9	56.4	78.8	86.3	1.0	7.3
Ours_E2E (ViT-L/14)	**57.4**	**79.8**	**87.8**	**1.0**	**9.2**	**58.9**	**80.7**	**89.5**	**1.0**	**7.2**

Table 5 Retrieval results on ActivityNet dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
T2VLAD	23.7	55.5	-	4.0	-	24.1	56.6	-	4.0	-
CLIP4Clip	41.4	73.7	85.3	2.0	-	-	-	-	-	-
VCM	40.8	72.8	-	2.0	7.3	42.6	74.9	86.2	2.0	6.4
CAMOE	51.0	77.7	-	-	-	49.9	71.4	-	-	-
DCR	46.2	77.3	88.2	2.0	-	45.7	76.5	87.8	2.0	-
Frozen	28.8	60.9	-	+ 3.0	-	-	-	-	-	-
Hun Yuan_tvr (ViT-B/16)	57.3	**84.8**	**93.1**	1.0	**4.0**	57.7	**85.7**	**93.9**	1.0	**3.4**
Hun Yuan_tvr (ViT-L/14)	55.3	83.3	92.2	1.0	4.3	55.8	84.1	93.2	1.0	3.6
MMT (ViT-B/16)	56.6	84.1	92.3	1.0	3.7	56.1	84.8	93.2	1.0	5.1
MMT (ViT-L/14)	55.7	84.3	92.8	1.0	5.2	57.3	85.1	93.5	1.0	5.0
Ours_Ensemble (ViT-B/16)	56.2	81.3	89.5	1.0	5.2	55.1	80.6	89.2	1.0	5.1
Ours_Ensemble (ViT-L/14)	**57.3**	81.9	89.4	**1.0**	5.8	**58.0**	82.6	90.7	**1.0**	4.6
Ours_E2E (ViT-B/16)	57.1	82.0	90.5	1.0	4.8	57.0	81.8	90.2	1.0	4.7
Ours_E2E (ViT-L/14)	56.4	81.1	89.1	1.0	5.9	57.4	82.2	90.6	1.0	4.6



A baby is standing in a kitchen. He is playing a harmonica. He is swirling back and forth dancing.

Do it again. Yeah. Job is. A? Oh, sp a harmonica. Yeah, you do a nice job. Careful, careful, wo multitasking here he's spinning in with harmonica. He's going to go over. Hold on. I do one minute.

A little boy pours mouthwash into his mouth from a cup. He spits it into the sink, then starts crying from the bad taste. He walks out of the room in tears.

Rira. So. Okay. That taste. A spicy. Is the best mouthwash ever? But. You have some water. Oh, god, i'm sorry. Now. A? Oh! Let you stay out, charlie. Look at one sport here, just drop some more. You are insane. And they?

Fig. 3 Qualitative comparisons of our method with CLIP4Clip and MMT. "Blue", "Green", "Orange", and "Purple" represent visual, speech, motion, and audio cues, respectively. "Red" means false retrieval results.

4.1.3 Pre-trained models

For MSR-VTT, LSMDC, and ActivityNet datasets, we directly utilize the shared audio and motion features from MMT, whose motion features are extracted using S3D trained on Kinetics [8] action recognition dataset and audio features are extracted using VGGish model [23] trained on YT8M. However, MMT does not provide multi-modal features for MSVD and DiDeMo datasets. Therefore, we extracted motion features using S3D trained on Kinetics and audio features using VGGish model [23] trained on AudioSet [19] for MSVD and DiDeMo datasets. Visual features are all extracted from ViT model of CLIP. Texts are all extracted using conformer model [22] trained on librispeech [37] and gigaspeech [10] datasets.

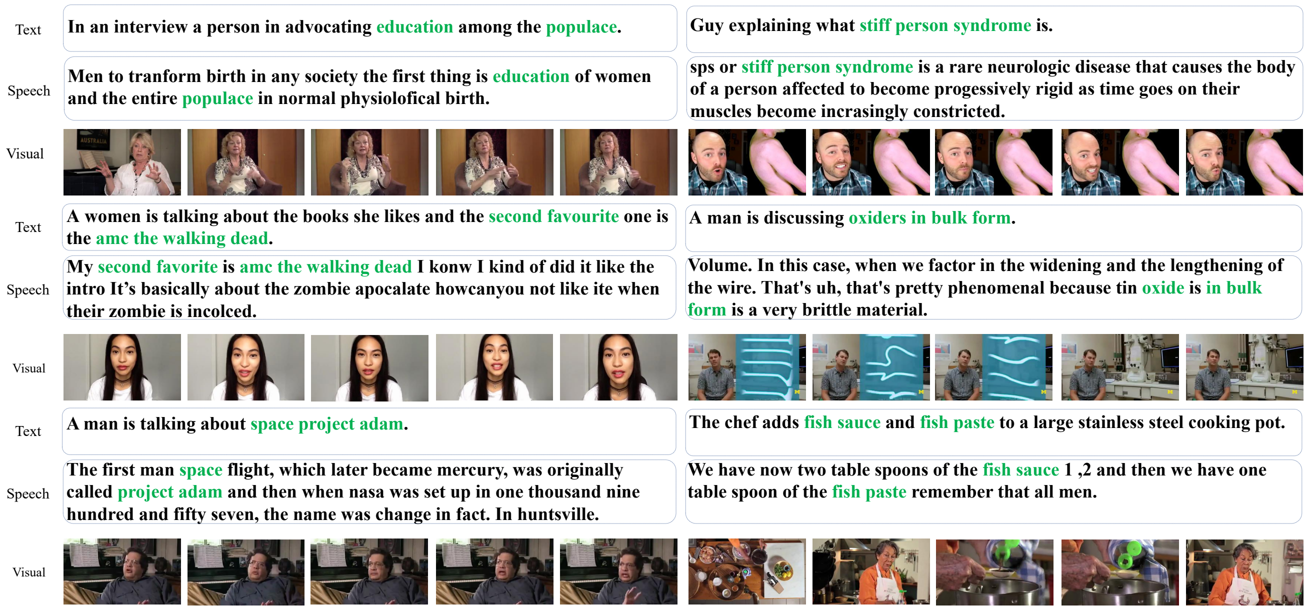
4.1.4 Implementation Details

Our method is implemented with PyTorch 1.7.1, and is trained on NVIDIA Tesla A100 GPU. All the experiments are trained on 8 GPUs. We set the initial learning rate as $1e-7$ for the CLIP and $1e-4$ for the remaining parameters, respectively, the temperature hyper-parameter λ in DSL is $1e-3$. The pre-trained weights in CLIP architectures including ViT-B/16 and ViT-L/14 are included in our work. Due to the big size of the above basic models, the sizes of the batch, frame length, and word length are also adjusted to adapt them. For MSR-VTT (ViT-B/16), MSVD (ViT-B/16), and LSMDC (ViT-B/16), the frame number F and token number T are 12 and 32, respectively, Adam optimizer with batch size (B) of 128 is used for training the model with 5 epoch. For DiDeMo (ViT-B/16) and ActivityNet (ViT-B/16), $F = 64$, $T = 64$, and $B = 32$. For MSR-VTT (ViT-L/14), MSVD (ViT-L/14), and LSMDC (ViT-L/14), $F = 32$, $T = 32$, and $B = 32$. For DiDeMo (ViT-L/14) and ActivityNet (ViT-L/14), $F = 64$, $T = 32$, and $B = 16$.

4.2 Comparison with State-of-the-art Methods

In this subsection, we compare M2HF with state-of-the-art methods, including T2VLAD [50], CLIP4Clip, VCM [7], CAMOE [12], X-Pool [21], LAFF [24], TVMM [30], DCR [46], Frozen [5], Hun Yuan_tvr, and MMT, on MSR-VTT, MSVD, LSMDC, DiDeMo, and ActivityNet benchmarks.

Table 1 shows results of MSR-VTT, which can be seen that our M2HF significantly surpasses CLIP4Clip by 21.5% R@1 and outperforms a very recent parallel work Hun Yuan_tvr by 11.0%. Table 2 shows that M2HF achieves 10.8% improvement on the MSVD compared to Hun Yuan_tvr. For LSMDC as shown in Table 3, our approach obtains the gain



In an interview a person in advocating education among the populace.

Men to tranform birth in any society the first thing is education of women and the entire populace in normal physiolofical birth.

Guy explaining what stiff person syndrome is.

sps or stiff person syndrome is a rare neurologic disease that causes the body of a person affected to become progessively rigid as time goes on their muscles become incrasingly constricted.

A women is talking about the books she likes and the second favourite one is the amc the walking dead.

My second favorite is amc the walking dead I konw I kind of did it like the intro It's basically about the zombie apocalate howcanyou not like ite when their zombie is incolced.

A man is discussing oxiders in bulk form.

Volume. In this case, when we factor in the widening and the lengthening of the wire. That's uh, that's pretty phenomenal because tin oxide is in bulk form is a very brittle material.

A man is talking about space project adam.

The first man space flight, which later became mercury, was originally called project adam and then when nasa was set up in one thousand nine hundred and fifty seven, the name was change in fact. In huntsville.

The chef adds fish sauce and fish paste to a large stainless steel cooking pot.

We have now two table spoons of the fish sauce 1 ,2 and then we have one table spoon of the fish paste remember that all men.

Fig. 4 Ablation studies of "Speech" branch in M2HF. The above examples all use the contents from the "Speech" branch to match the correct text. "Speech" branch can provide very abstract contents, such as "populace", "space project adan", and "stiff person syndrome", which cannot be provided by other modalities.

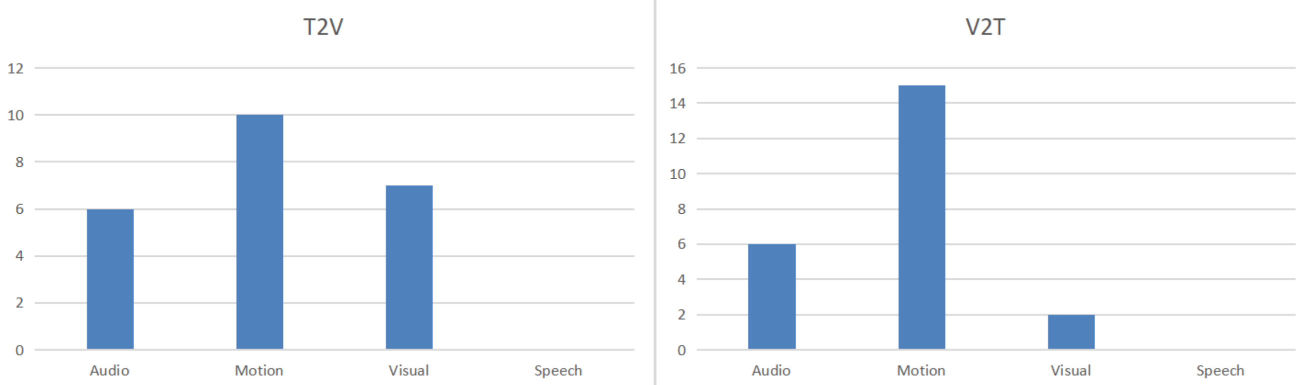
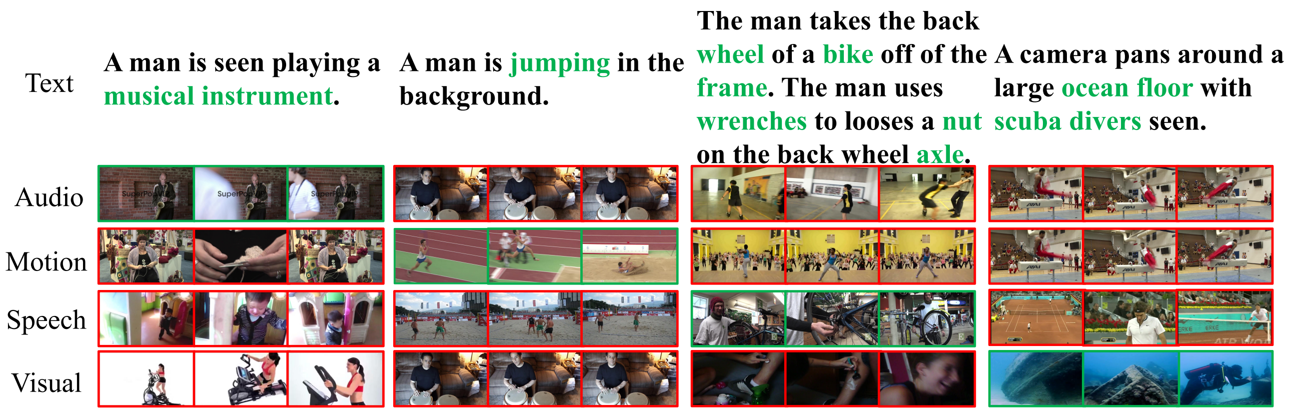


Fig. 5 The frequency statistics of each branch with the greatest contribution. A dataset has 4 groups (Ours_Ensemble (ViT-B/16), Ours_Ensemble (ViT-L/14), Ours_E2E (ViT-B/16), and Ours_E2E (ViT-L/14)) shown in the Table 6-Table 10. We count the frequency of each branch being the greatest contribution from a total of 20 groups for five datasets.



A man is seen playing a musical instrument.

A man is jumping in the background.

The man takes the back wheel of a bike off of the frame. The man uses wrenches to looses a nut on the back wheel axle.

A camera pans around a large ocean floor with scuba divers seen.

Fig. 6 Ablation studies of M2HF. "Audio", "Motion", "Speech", and "Visual" are the model of only using corresponding branch.

Table 6 Ablation studies on MSR-VTT 1K dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Audio (ViT-B/16)	52.8	78.0	86.1	1.0	10.3	52.9	77.3	86.4	1.0	8.0
Motion (ViT-B/16)	53.4	78.2	86.2	1.0	9.9	53.0	78.3	86.5	1.0	8.3
Speech (ViT-B/16)	5.7	12.5	14.1	-	-	5.6	11.1	12.3	-	-
Visual (ViT-B/16)	53.2	79.1	86.9	1.0	9.9	52.7	78.5	86.3	1.0	7.3
Audio (ViT-L/14)	53.0	77.6	85.3	1.0	11.7	54.7	77.5	85.1	1.0	8.6
Motion (ViT-L/14)	53.3	77.1	85.7	1.0	11.4	53.2	77.1	86.3	1.0	8.6
Speech (ViT-L/14)	5.7	12.5	14.1	-	-	5.6	11.1	12.3	-	-
Visual (ViT-L/14)	53.2	77.6	85.8	1.0	11.9	53.2	76.5	85.1	1.0	9.1
w/o Audio	59.3	83.1	89.7	1.0	8.1	60.4	83.2	90.2	1.0	5.7
w/o Motion	59.1	83.0	89.8	1.0	8.1	59.5	82.9	89.6	1.0	5.7
w/o Speech	59.2	82.4	88.9	1.0	8.0	60.6	82.8	90.4	1.0	5.4
w/o Visual	58.9	82.5	89.3	1.0	8.2	59.6	82.7	89.9	1.0	6.0
Ours_Ensemble (ViT-B/16)	59.9	82.8	89.3	1.0	8.1	60.7	82.7	90.2	1.0	5.6
w/o Audio	61.8	83.4	90.1	1.0	8.5	61.8	83.7	90.5	1.0	5.7
w/o Motion	62.5	83.7	90.6	1.0	8.7	63.5	83.9	90.6	1.0	5.7
w/o Speech	63.8	84.2	90.6	1.0	8.3	64.5	84.7	90.8	1.0	5.4
w/o Visual	62.3	82.5	90.2	1.0	9.0	62.6	83.9	90.5	1.0	5.9
Ours_Ensemble (ViT-L/14)	65.3	85.3	91.7	1.0	7.9	66.0	86.0	91.7	1.0	5.0
w/o Audio	59.4	83.1	89.2	1.0	7.5	59.4	84.1	90.1	1.0	6.3
w/o Motion	59.5	83.5	89.1	1.0	8.3	58.8	83.0	89.8	1.0	6.2
w/o Speech	60.6	83.8	89.8	1.0	6.9	61.9	84.6	91.0	1.0	5.5
w/o Visual	59.9	83.3	89.7	1.0	7.5	59.8	82.6	90.2	1.0	6.2
Ours_E2E (ViT-B/16)	62.1	84.8	90.7	1.0	6.5	63.5	85.6	91.6	1.0	5.6
w/o Audio	61.3	83.3	89.6	1.0	8.3	60.6	82.3	89.9	1.0	6.4
w/o Motion	62.4	84.0	90.3	1.0	8.7	62.3	84.4	90.7	1.0	5.8
w/o Speech	64.5	85.2	90.5	1.0	7.0	64.3	85.1	91.1	1.0	5.4
w/o Visual	61.6	82.2	89.4	1.0	8.2	62.8	83.1	90.1	1.0	6.6
Ours_E2E (ViT-L/14)	66.0	86.3	91.5	1.0	6.7	65.7	86.2	91.9	1.0	5.1

over Hun Yuan_tvr by 4.2%. As reported in Table 4, M2HF remarkably outperforms the state-of-the-art method by 5.3% for DiDeMo. Table 5 demonstrates comparable improvement performance on ActivityNet compared to Hun Yuan_tvr by 2.2% on V2T task. To maintain the fairness of the experiment, the multi-modal features fed into MMT are the same as M2HF. According to Table 1 to Table 5, we can find that compared with the MMT, M2HF significantly outperforms 10.1%, 8.2%, 3.5%, 4.2%, and 0.7% for MSRVTT, MSVD, LSMDC, DiDeMo, and ActivityNet, respectively. M2HF can better fuse multi-modal features when some of these features are weak and some are powerful. All the quantitative results consistently illustrate the superiority of M2HF.

Fig. 3 shows two qualitative comparison examples, which show that only using visual modality is not enough to represent videos well. In contrast, our multi-modal complement method provides multi-modal cues to obtain more accurate results. For the first example, images are helpful in matching the "baby" word. The harmonica sound made by this baby and the text information from off-screen sound can be associated with "harmonica". The baby's movements are matched with "swirling back" and "forth dancing". The result retrieved via the visual-based method CLIP4clip is completely unrelated to the text. The result retrieved via MMT can only retrieve "baby". In the second example, "little boy" corresponds to a semantic visual target. The keywords "mouthwash" and "taste" in the text match the relevant tokens in the caption text. The "crying" sound made by this little boy is captured with the help of audio signals. His moving figures are related to "walk out of". However, the visual-based method CLIP4clip and MMT can only retrieve "a little boy".

M2HF pioneers the use of a more interpretable early-fusion technique in video retrieval tasks and the use of a better late fusion way to fuse different-branch components, which significantly improves the performance.

4.3 Ablation Studies

Effect of multi-branch strategy. As reported in Table 6 - Table10, detailed ablation studies for each dataset are performed to prove that every modality of each branch contributes to retrieving the correct results. Specifically, "Audio", "Motion", "Speech", and "Visual" are the performance using only the respective branch. And "w/o Audio", "w/o Motion", "w/o Speech", and "w/o Visual" are the performance of models

Table 7 Ablation studies on MSVD dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Audio (ViT-B/16)	55.0	82.4	89.4	1.0	7.9	61.3	80.8	90.2	1.0	4.7
Motion (ViT-B/16)	55.4	82.7	89.4	1.0	7.8	61.2	82.7	90.7	1.0	5.3
Visual (ViT-B/16)	54.9	82.4	89.6	1.0	7.8	63.1	82.0	90.4	1.0	5.3
Audio (ViT-L/14)	58.2	83.6	89.9	1.0	7.6	61.8	83.6	91.8	1.0	3.8
Motion (ViT-L/14)	58.3	83.3	89.4	1.0	7.8	64.3	85.8	91.2	1.0	3.9
Visual (ViT-L/14)	57.2	82.7	89.2	1.0	8.1	59.4	80.4	89.6	1.0	4.6
w/o Audio	59.9	85.2	91.0	1.0	6.7	67.8	86.9	93.6	1.0	3.8
w/o Motion	58.7	84.6	90.6	1.0	7.0	65.8	84.5	91.9	1.0	3.9
w/o Visual	59.9	85.1	90.9	1.0	6.7	67.2	85.7	93.3	1.0	3.6
Ours_Ensemble (ViT-B/16)	61.7	86.1	91.5	1.0	6.3	69.1	87.3	93.9	1.0	3.4
w/o Audio	63.9	86.4	91.6	1.0	6.2	68.2	87.6	94.5	1.0	3.0
w/o Motion	63.9	86.4	91.7	1.0	6.3	67.9	86.3	94.0	1.0	3.1
w/o Visual	64.4	86.9	91.9	1.0	6.0	70.1	88.4	94.5	1.0	2.8
Ours_Ensemble (ViT-L/14)	67.1	88.1	92.7	1.0	5.5	72.1	89.3	95.5	1.0	2.6
w/o Audio	60.2	85.1	91.1	1.0	6.6	66.7	85.4	94.0	1.0	3.1
w/o Motion	59.3	84.6	90.8	1.0	6.8	66.7	85.4	94.0	1.0	3.1
w/o Visual	60.7	85.2	91.0	1.0	6.5	70.7	89.3	93.6	1.0	2.9
Ours_E2E (ViT-B/16)	62.7	86.3	91.7	1.0	6.0	73.3	90.6	94.3	1.0	2.7
w/o Audio	64.7	86.9	91.6	1.0	5.9	72.1	90.4	94.5	1.0	3.0
w/o Motion	64.5	86.7	91.7	1.0	7.0	65.8	85.7	93.3	1.0	3.5
w/o Visual	64.5	86.8	91.9	1.0	6.0	75.7	91.3	96.0	1.0	2.5
Ours_E2E (ViT-L/14)	68.6	88.7	92.9	1.0	5.1	75.7	91.0	96.3	1.0	2.6

using all branches except for the relevant branch. There is no “Speech” branch in the MSVD dataset since there are no audio signals for those videos. Moreover, its “Audio” branch works depended on our proposed multi-modal completion strategy.

To model the relationship between video features and text features, some simple and direct fusion methods can be applied to fuse multi-modal features from video, which may result in multi-modal confusion. The superiority of our multi-branch way is shown in this paper. Table 11 has explored th effect of multi-branch strategy on MSR-VTT 1K dataset. The general one-branch way is fusing all multi-modal features to obtain a joint video representation, which is then used to make a relationship with text features. Three common approaches are explored including “Multiply”, “Average”, and “Add”. As reported in Table 11, video features, which are obtained in these ways, to make connections with text features remarkably decrease performance compared to our proposed “Multi-branch” way. The reason is that the way to fuse all features yields summarized features, which may result in multi-modal confusion. Our proposed method can make full use of every modality feature.

Only one branch. For the case of using only one branch (see the top two groups of Table 6 -Table10), we can see that the performance of the “Speech” branch is poorer than that of other branches. This is attributed to its absence from many videos and the limited effect of the Jaccard method, and exploring a better method for the “Speech” branch is left to our future direction. However, text cues from videos and caption text queries are the same modality, which means that their relationships are most intimate. Furthermore, this branch can provide abundant abstract information for video retrieval, which cannot be provided by other branches. Some representative examples of the “Speech” branch are shown in Fig. 4, the visual images of these examples are people talking about something very abstract, such as “populace”, “oxide”, “space project adan”, and “stiff person syndrome”, which cannot understand from other modalities. This phenomenon is common for text-video retrieval tasks. Compared to only using the “Visual” branch, only using the “Audio” and “Motion” branches can also obtain similar even better performance, which implies that the guidance mechanism provides different perspectives such as sound and moving objects for more fine-grained retrieval results.

Without one branch. Other experiments in Table 6-Table 10 are ablation studies removing “Audio”, “Motion”, “Speech”, and “Visual” branches to verify their impact on the performance of relevant models. First, we can see that all these modalities make their contributions to improving the performance of final retrieval results. Then, different datasets prefer different modalities since the focus of each dataset may be different. However, we are still interested in figuring out which modality makes the greatest contribution to the T2V and V2T task tasks. Therefore, we conduct a statistical analysis as reported in Fig. 5. For each group experiment in Table 6-Table 10, *e.g.*, “w/o Audio”, “w/o

Table 8 Ablation studies on LSMDC dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Audio (ViT-B/16)	21.6	39.0	48.7	11.0	63.2	22.1	39.4	48.4	12.0	57.2
Motion (ViT-B/16)	22.7	40.0	49.3	11.0	60.8	23.2	40.6	48.8	12.0	55.4
Speech (ViT-B/16)	0.6	0.8	0.8	-	-	0.2	0.5	0.7	-	-
Visual (ViT-B/16)	23.8	40.6	51.2	10.0	59.8	23.0	41.6	51.1	10.0	50.9
Audio (ViT-L/14)	23.4	42.9	51.6	9.0	58.3	22.8	43.8	52.3	9.0	49.6
Motion (ViT-L/14)	24.8	44.7	53.0	8.0	53.0	26.3	43.8	53.8	8.0	45.8
Speech (ViT-L/14)	0.6	0.8	0.8	-	-	0.2	0.5	0.7	-	-
Visual (ViT-L/14)	25.8	45.4	54.8	7.0	54.7	25.4	47.5	55.5	7.0	44.9
w/o Audio	28.7	45.4	56.7	7.0	46.8	28.3	46.8	56.3	7.0	39.9
w/o Motion	27.3	45.1	55.8	7.5	49.0	26.9	45.6	55.6	7.0	42.0
w/o Speech	30.2	46.8	58.2	7.0	43.2	29.8	48.1	57.8	6.0	37.2
w/o Visual	26.8	43.5	53.7	9.0	51.0	27.4	44.2	53.3	8.0	46.3
Ours_Ensemble (ViT-B/16)	30.2	46.8	58.2	7.0	43.2	29.8	48.1	57.8	6.0	37.2
w/o Audio	32.0	53.3	61.8	5.0	38.5	31.9	54.3	62.6	4.0	31.9
w/o Motion	30.5	51.5	60.4	5.0	42.8	30.1	54.0	61.6	4.0	33.7
w/o Speech	33.6	55.8	64.1	4.0	34.8	33.9	56.9	64.5	3.0	28.4
w/o Visual	29.4	50.8	59.2	5.0	41.4	30.2	51.0	59.3	5.0	35.9
Ours_Ensemble (ViT-L/14)	33.9	55.9	64.2	4.0	34.8	34.0	57.1	64.6	3.0	28.3
w/o Audio	25.3	42.7	52.1	9.0	53.5	27.0	44.2	53.1	9.0	46.3
w/o Motion	25.0	42.6	50.5	10.0	55.0	24.0	41.5	50.4	10.0	50.7
w/o Speech	27.1	44.9	53.0	9.0	46.7	26.9	45.5	53.6	8.0	43.2
w/o Visual	26.7	42.9	51.9	9.0	53.5	27.0	43.2	53.2	8.0	46.3
Ours_E2E(ViT-B/16)	30.3	46.4	55.9	7.0	44.8	29.1	46.8	56.3	6.0	39.6
w/o Audio	28.3	47.9	57.3	6.0	47.3	29.5	48.1	58.5	6.0	41.3
w/o Motion	27.7	45.6	56.3	7.0	46.9	28.4	45.8	54.4	8.0	46.3
w/o Speech	31.1	48.8	58.4	6.0	44.5	30.4	48.4	59.4	6.0	38.1
w/o Visual	27.2	45.0	56.1	7.0	49.8	28.7	47.7	57.1	6.0	41.5
Ours_E2E (ViT-L/14)	31.8	53.2	62.2	4.0	36.9	31.6	53.3	63.6	5.0	30.8

Motion", "w/o Speech", "w/o Visual", and "Ours_E2E (ViT-B/16)" in Table 6-Table 10, we record the relevant branch that contributes the most according to the rule with the most performance (R@1) degradation compared to the other branches. Specifically, we count the frequency of each branch being the greatest contribution from a total of 20 groups for five datasets. For example, in Table 6, the 9th - the 13th lines represent Ours_Ensemble (ViT-B/16) group. The performance of Ours_Ensemble (ViT-B/16) is 59.9%, however, the worst ablation branch performance is "w/o Visual" (58.9%). Therefore, the visual branch has the most contribution to this group. The frequency for branch "Visual" will add 1. As shown in Fig. 5, the "Motion" branch provides the greatest contribution in either T2V or V2T task. The reason might be that motion features are available for all examples and the guidance of motion features can surely capture the moving objects matching the verb phrase in text queries. "Visual" and "Audio" branchs are in the second place for T2V task and V2T task respectively. This is the same as the fact that human perception is most dependent on visual and audio perceptions. The poor performance of "Speech" is attributed to its absence from many videos, but its contributions cannot be ignored. To summarize, multi-modal perception is indeed more reasonable and effective than only using one modality.

Effect of multi-modal strategy. As reported in Table 6-Table 10, ablation experiments are conducted to evaluate the effect of each modality in M2HF. "w/o Audio", "w/o Motion", "w/o Speech", and "w/o Visual" represents removing the relevant modality from M2HF. Quantitative results demonstrate that each modality contributes to the performance of text-video retrieval. Fig. 6 shows the qualitative studies of our proposed method. "Audio", "Motion", "Speech", and "Visual" are the effect of only using corresponding modality for T2V retrieval. The green and red boxes represent the true and false retrieval results, respectively. These four examples explain the advantages of each modality. For the first column, the "Audio" branch can catch the instrument's sound source with the guidance of an audio signal, however, other modalities cannot provide the same contribution. The second example utilizes the motion features of jumping to predict the right retrieval, where the "Motion" branch can pay attention to the moving objects. The third one shows the effect of the "Speech" branch, and there are six same keywords between caption text tokens and speech text tokens, including "wheel",

Table 9 Ablation studies on DiDeMo dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Audio (ViT-B/16)	46.6	74.8	82.2	2.0	13.5	49.8	73.8	82.8	2.0	9.7
Motion (ViT-B/16)	47.0	73.9	81.3	2.0	13.4	46.3	73.1	82.0	2.0	10.1
Speech (ViT-B/16)	1.9	3.2	3.6	-	-	0.4	1.3	1.7	-	-
Visual (ViT-B/16)	46.4	74.4	81.4	2.0	13.4	47.0	73.2	82.7	2.0	9.9
Audio (ViT-L/14)	48.3	75.0	83.3	2.0	12.6	50.7	76.2	84.2	1.0	9.3
Motion (ViT-L/14)	47.9	74.0	82.2	2.0	12.4	48.6	74.6	84.2	2.0	9.5
Speech (ViT-L/14)	1.9	3.2	3.6	-	-	0.4	1.3	1.7	-	-
Visual (ViT-L/14)	48.9	73.7	82.5	2.0	12.9	49.1	74.9	84.8	2.0	9.7
w/o Audio	53.1	78.3	84.4	1.0	10.9	52.4	77.0	85.1	1.0	7.9
w/o Motion	52.8	78.1	84.3	1.0	11.2	53.5	77.2	84.9	1.0	8.3
w/o Speech	54.8	79.5	85.3	1.0	10.0	56.0	79.0	86.1	1.0	7.3
w/o Visual	52.3	78.0	84.3	1.0	10.9	53.7	77.4	84.7	1.0	8.1
Ours_Ensemble (ViT-B/16)	54.8	79.5	85.3	1.0	10.0	56.0	79.0	86.1	1.0	7.3
w/o Audio	55.2	77.4	85.3	1.0	10.6	55.9	78.4	87.4	1.0	7.9
w/o Motion	54.1	77.8	85.8	1.0	10.9	55.8	79.0	87.5	1.0	8.0
w/o Speech	56.8	79.6	87.2	1.0	9.6	58.4	80.6	88.8	1.0	7.3
w/o Visual	54.5	78.3	86.3	1.0	10.4	55.7	79.2	87.2	1.0	7.8
Ours_Ensemble (ViT-L/14)	57.1	79.6	87.3	1.0	9.6	58.4	80.6	88.8	1.0	7.3
w/o Audio	54.5	78.0	84.4	1.0	10.9	52.8	77.0	85.5	1.0	8.0
w/o Motion	52.9	77.8	84.1	1.0	11.2	53.6	77.2	85.2 -	1.0	8.3
w/o Speech	56.3	79.2	85.1	1.0	10.0	56.4	78.8	86.3	1.0	7.3
w/o Visual	53.7	77.8	83.8	1.0	10.8	54.3	77.4	85.2	1.0	8.1
Ours_E2E (ViT-B/16)	56.9	79.3	85.3	1.0	9.9	56.4	78.8	86.3	1.0	7.3
w/o Audio	55.9	78.0	86.1	1.0	10.1	56.0	78.6	88.1	1.0	7.8
w/o Motion	54.1	77.7	85.9	1.0	10.8	55.4	78.9	87.5	1.0	8.0
w/o Speech	57.1	79.8	87.8	1.0	9.2	58.9	80.7	89.5	1.0	7.2
w/o Visual	54.5	77.9	86.6	1.0	9.9	56.4	79.2	87.7	1.0	7.7
Ours_E2E (ViT-L/14)	57.4	79.8	87.8	1.0	9.2	58.9	80.7	89.5	1.0	7.2

"bike", "frame", "wrench", "nut", and "axle". This branch is really helpful for those cases containing lots of abstract nouns. For the last example, there are no sound and moving objects. Therefore, "ocean floor" and "scuba divers" visual contents are detected by the "Visual" branch. To this end, quantitative and qualitative results illustrate the superiority of multi-modal for TVR.

Effect of early fusion. Instead of directly using audio and motion features to compute the similarity with text features, in this work, we apply early fusion to fuse these two kinds of features with visual features respectively, as shown in the left bottom of Fig. 2. The starting point is that there exists a large gap in representation ability between visual features extracted by CLIP and audio and motion features extracted by relatively weak models, and powerful visual features are efficiently augmented through this guidance-based fusion for focusing different aspects with the help of other modal features.

To evaluate the effectiveness of early fusion, an ablation study is conducted in Table 12 on MSR-VTT 1k dataset. From this table, we can find that the retrieval result of directly using original audio and motion features without early fusion is 59.4%, which is significantly lower than that of the proposed method (66.0%). The results prove that early fusion can improve motion features and audio features and directly using audio and motion features to compute the similarity with text features will harm the performance of the relative branch. The reason is that weak motion features and audio features extracted by relatively weak pre-trained models fuse with powerful visual features extracted by powerful CLIP are efficiently augmented through this guidance-based fusion for focusing on the moving objects and the content of making sounds. Early fusion can effectively exploit the comprehensiveness of multi-modal features to achieve better performance. Using equally powerful multi-modal features will be explored in future work.

Effect of fusion strategy of early fusion. The effect of early fusion strategies for each branch is also worth studying. As shown in Table 13, we explore the effect of the fusion strategy of early fusion on MSR-VTT 1k dataset. "AM_MD" represents that early fusion for "Audio" and "Motion" are MFB and dual transformer respectively, and this setting obtains the best performance (66.0%). Three other settings include that early fusion for "Audio" and "Motion" both are MFB, which is denoted as "AM_MM" (65.8%), early fusion for "Audio"

Table 10 Ablation studies on ActivityNet dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Audio (ViT-B/16)	46.0	74.5	84.4	2.0	7.7	45.5	73.2	83.9	2.0	7.8
Motion (ViT-B/16)	46.1	74.4	84.2	2.0	7.8	45.8	73.3	83.8	2.0	8.2
Speech (ViT-B/16)	2.3	5.1	7.8	-	-	3.4	7.1	9.8	-	-
Visual (ViT-B/16)	46.6	74.5	84.7	2.0	7.4	45.6	73.0	83.7	2.0	7.7
Audio (ViT-L/14)	47.1	74.2	84.4	2.0	8.4	47.1	75.2	85.4	2.0	6.7
Motion (ViT-L/14)	47.5	75.8	85.0	2.0	8.1	48.4	76.3	86.1	2.0	7.2
Speech (ViT-L/14)	2.3	5.1	7.8	-	-	3.4	7.1	9.8	-	-
Visual (ViT-L/14)	45.7	73.2	83.1	2.0	9.3	46.0	73.0	83.9	2.0	7.6
w/o Audio	54.7	80.3	89.0	1.0	5.2	54.4	80.3	89.1	1.0	5.1
w/o Motion	53.7	80.2	88.6	1.0	5.4	53.0	79.1	88.4	1.0	5.4
w/o Speech	56.2	81.3	89.5	1.0	5.2	55.1	80.6	89.2	1.0	5.1
w/o Visual	54.3	80.7	88.8	1.0	5.5	53.8	79.9	88.8	1.0	5.4
Ours_Ensemble (ViT-B/16)	56.2	81.3	89.5	1.0	5.2	55.1	80.6	89.2	1.0	5.1
w/o Audio	55.0	80.7	88.6	1.0	6.0	56.1	81.3	89.9	1.0	4.8
w/o Motion	51.7	78.0	86.8	1.0	7.3	52.5	78.4	87.8	1.0	5.8
w/o Speech	56.0	80.8	88.5	1.0	6.3	56.3	81.3	89.7	1.0	5.0
w/o Visual	54.5	80.6	88.4	1.0	6.3	55.8	81.3	89.8	1.0	4.9
Ours_Ensemble (ViT-L/14)	57.3	81.9	89.4	1.0	5.8	58.0	82.6	90.7	1.0	4.6
w/o Audio	54.6	80.4	89.4	1.0	5.3	54.3	79.8	89.0	1.0	5,2
w/o Motion	53.1	79.6	88.5	1.0	5.5	52.7	79.0	88.3	1.0	5.5
w/o Speech	55.8	81.0	89.6	1.0	5.2	55.2	80.6	89.0	1.0	5.2
w/o Visual	54.0	80.5	89.1	1.0	5.4	54.0	79.8	89.0	1.0	5.3
Ours_E2E (ViT-B/16)	57.1	82.0	90.5	1.0	4.8	57.0	81.8	90.2	1.0	4.7
w/o Audio	51.9	76.9	85.0	1.0	11.1	53.7	78.2	87.0	1.0	7.8
w/o Motion	51.0	75.9	84.9	1.0	12.3	52.9	76.5	86.6	1.0	8.6
w/o Speech	53.3	77.6	84.5	1.0	11.1	52.5	77.3	84.5	1.0	11.1
w/o Visual	53.4	79.1	85.6	1.0	10.8	52.1	77.5	87.7	1.0	8.3
Ours_E2E (ViT-L/14)	56.4	81.1	89.1	1.0	5.9	57.4	82.2	90.6	1.0	4.6

Table 11 Effect of multi-branch strategy on MSR-VTT 1K dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
M2HF_multiply	49.1	74.7	84.7	2.0	13.0	49.0	75.0	84.1	2.0	10.2
M2HF_average	52.9	77.3	85.7	1.0	11.4	53.0	78.0	86.3	1.0	8.6
M2HF_add	50.7	76.6	85.2	1.0	12.7	52.1	76.9	85.3	1.0	11.4
M2HF_multi_branch	66.0	86.3	91.5	1.0	6.7	65.7	86.2	91.9	1.0	5.1

Table 12 Effect of early fusion on MSR-VTT 1K dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
without_early_fusion	59.4	82.7	89.5	1.0	7.6	59.4	82.7	90.0	1.0	5.8
with_early_fusion	66.0	86.3	91.5	1.0	6.7	65.7	86.2	91.9	1.0	5.1

Table 13 Effect of fusion strategy of early fusion on MSR-VTT 1K dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
AM_MM	65.8	86.8	92.0	1.0	5.7	65.5	87.2	91.6	1.0	4.7
AD_MM	66.0	85.9	91.8	1.0	6.7	65.4	86.8	92.0	1.0	5.0
AD_MD	65.6	85.4	91.3	1.0	7.3	65.2	85.7	91.7	1.0	5.5
AM_MD	66.0	86.3	91.5	1.0	6.7	65.7	86.2	91.9	1.0	5.1

Table 14 Ablation studies of MMBL on MSR-VTT 1K dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Average	65.5	84.9	91.3	1.0	7.8	64.4	86.1	92.0	1.0	5.5
Element-wise adding	65.8	84.8	91.2	1.0	7.8	64.5	85.9	91.6	1.0	5.5
Element-wise maximizing	65.3	84.7	91.1	1.0	8.2	64.6	86.4	91.5	1.0	5.5
Element-wise minimizing	66.0	86.3	91.5	1.0	6.7	65.7	86.2	91.9	1.0	5.1

Table 15 Effect of multi-modal completion on DiDeMo dataset.

Methods	T2V					V2T				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Without_Completion	56.5	78.9	87.1	1.0	10.0	57.7	79.9	88.1	1.0	7.3
With_Completion	57.4	79.8	87.8	1.0	9.2	58.9	80.7	89.5	1.0	7.2

and "Motion" both are dual transformer, which is denoted as "AD_MD" (65.6%), and early fusion for "Audio" and "Motion" are dual transformer and MFB respectively, which is denoted as "AD_MM" (66.0%). We can see that different early fusion settings obtain similarly improved performance in Table 13. The subtle differences identify the robustness of different early fusion strategies for different branches. However, they all achieve better performance compared to the setting without early fusion (59.4%) in the first row of Table 12. Why does using MFB for audio-visual fusion and using a dual transformer for animation fusion achieve better results? The reason might be that the MFB strategy has a strong ability to fuse local audio-visual features to obtain local fusion features. However, the dual transformer strategy is based on a self-attention mechanism. It has strong modeling capabilities for moving objects among sequences. Taking time as the axis, audio-visual fusion has more local properties and motion-visual fusion has more global properties. In other words, audio features are more relevant to images in nearby frames and less relevant to further frames, such as a barking dog. However, dual transformer architecture may damage the audio-visual fusion. Motion features are relevant to all visual frames, such as a moving car. However, MFB can only fuse the local motion-visual features, which may damage the performance.

Ablation study for MMBL. To explore the best way to late fuse all branches, an ablation study of MMBL on MSR-VTT 1K dataset is conducted, and the numerical results are reported in Table 14 for the E2E training. MMBL is a multi-modal balance loss for training the model in an end-to-end manner. There are at least four ways to late fuse all branches, including taking the average value (65.5%), taking the element-wise maximum value (65.3%), taking the element-addition value (65.8%), and taking the element-wise minimum value (66.0%). Finally, we introduce a multi-modal balance loss by taking the element-wise minimum of each branch according to its superior performance.

Effect of multi-modal completion. Due to the existence of modality absence, a multi-modal completion method is proposed in our work. When removing the multi-modal completion (in Sec. 3.3), Table 15 illustrates the effect of multi-modal completion. The audio signal is not available to some samples in DiDeMo. Therefore, the multi-modal completion method is beneficial to mitigate modality absence. When removing the multi-modal completion, the retrieval performance drops from 57.4% to 56.5% on DiDeMo.

Analysis of computation cost. 8 A100 GPUs are used to train M2HF on all datasets. On MSRVTT, the training time of M2HF is 3.4 hours per epoch, and CLIP4Clip is 2.5 hours per epoch. The inference of each branch is independent and is optimized as a multi-branch parallel inference. Therefore, the inference time per video on an A100 GPU of M2HF (38.7ms) depends on the longest branch, which is not so much longer than CLIP4Clip (24.4ms). The average GPU memory cost is 27.5G, compared with CLIP4Clip (20.8G).

5 Conclusions

Based on the multi-modal nature of videos, in this paper, we proposed a novel multi-branch multi-modal hybrid fusion network for text-video retrieval. The core idea is to explore fine-grained multi-modal cues in a multi-branch way, and M2HF can also leverage the powerful knowledge from a pre-trained text-image retrieval model (*i.e.*, CLIP). Two training strategies are exploited and implemented: end-to-end training with a multi-modal balance loss and ensemble training with a multi-modal balance fusion. Extensive quantitative and qualitative comparisons and ablation experiments are conducted to validate our method. M2HF has achieved state-of-the-art performance for TVR on MSR-VTT, MSVD, LSMDC, DiDeMo, and ActivtiyNet. In the future, we would like to improve the multi-modality abilities with more features, such as optical flow.

Acknowledgements

5.1 Funding

This work was partially supported by the National Natural Science Foundation of China (62102418 and 62172415), the CCF-Tencent Rhino-Bird Research Fund (RAGR20210124), and the Beijing Science and Technology Plan Project (Z231100005923033).

5.2 Authors' contributions

Conceptualization, S.L., S.C., and K.Y.; methodology, S.L., S.C., and W.Q.; software, S.L., M.Z., C.C., and Z.Z.; validation, S.C., J.K., and M.Z.; formal analysis, S.L., W.Q., and Z.Z.; writing—original draft preparation, S.L., W.Q., and M.Z.; writing—review and editing, S.L., S.C., C.C., and D.Y.; visualization, J.K. and Z.Z.; supervision, K.Y., W.Q., and D.Y.;

Declaration of competing interest

The authors have no competing interests.

References

[1] S. Alghowinem, R. Göcke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear. Cross-cultural detection of depression from nonverbal behaviour. *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 1:1–8, 2015. 3

[2] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5803–5812, 2017. 7

[3] P. Atrey, M. Kankanhalli, and R. Jain. Information assimilation framework for event detection in multimedia surveillance systems. *Multimedia Systems*, 12:239–253, 12 2006. 3

[4] S. Ayache, G. Quénot, and J. Gensel. Classifier fusion for svm-based multimedia semantic indexing. In *European Conference on Information Retrieval*, 2007. 3

[5] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1708–1718, 2021. 4, 9

[6] A. Bendjebbour, Y. Delignon, L. Fouque, V. Samson, and W. Pieczynski. Multisensor image segmentation using dempster-shafer fusion in markov fields context. *IEEE Transactions on Geoscience and Remote Sensing*, 39:1789–1798, 2001. 2

[7] S. Cao, B. Wang, W. Zhang, and L. Ma. Visual consensus modeling for video-text retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 167–175, 2022. 1, 9

[8] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733, 2017. 9

[9] D. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of theannual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 7

[10] G. Chen, S. Chai, G.-B. Wang, J. Du, W. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Y. Wang, Z. You, and Z. Yan. Gigaspeech: An evolving, multi-domain asr corpus with 10, 000 hours of transcribed audio. In *Interspeech*, 2021. 9

[11] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, and H. Sahli. Efficient spatial temporal convolutional features for audiovisual continuous affect recognition. In *Proceedings of the International on Audio/Visual Emotion Challenge and Workshop*, 2019. 2

[12] X. Cheng, H. Lin, X. Wu, F. Yang, and D. Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 4, 7, 9

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019. 4

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations*, 2021. 4

[15] M. Dzabraev, M. Kalashnikov, S. Komkov, and A. Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3363, 2021. 1, 3

[16] H. Fang, P. Xiong, L. Xu, and Y. Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 4

[17] H. Fang, P. Xiong, L. Xu, and W. Luo. Transferring image-clip to video-text retrieval via temporal relations. *IEEE Transactions on Multimedia*, pages 1–14, 2022. 1

[18] V. Gabeur, C. Sun, K. Alahari, and C. Schmid. Multi-modal transformer for video retrieval. In *Proceedings of the European Conference on Computer Vision*, page 214–229, 2020. 1

[19] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780, 2017. 4, 9

[20] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011. 3

[21] S. K. Gorti, N. Vouitsis, J. Ma, K. Golestan, M. Volkovs, A. Garg, and G. Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5006–5015, 2022. 4, 9

[22] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. Conformer: Convolution-augmented transformer for speech recognition. *ArXiv*, 2020. 9

[23] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 131–135, 2017. 4, 9

[24] F. Hu, A. Chen, Z. Wang, F. Zhou, J. Dong, and X. Li. Lightweight attentional feature fusion: A new baseline for text-to-video retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 444–461, 2022. 4, 9

[25] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 2, 5, 6

[26] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. In *Proceedings of the International Conference on Learning Representations*, pages 247–263, 2017. 2

[27] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715, 2017. 7

[28] A. Kunitsyn, M. Kalashnikov, M. Dzabraev, and A. Ivaniuta. Mdmmt-2: Multidomain multimodal transformer for video retrieval, one more step towards generalization. *arXiv preprint arXiv:2203.07086*, 2022. 3

[29] Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. Hauptmann. Multimedia classification and event detection using double fusion. *Multimedia Tools and Applications*, 71, 07 2014. 3

[30] C. Lin, A. Wu, J. Liang, J. Zhang, W. Ge, W.-S. Zheng, and C. Shen. Text-adaptive multiple visual prototype matching for video-text retrieval. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2022. 4, 9

[31] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 7

[32] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 1

[33] Z. Ma, F. Ma, B. Sun, and S. Li. Hybrid mutimodal fusion for dimensional emotion recognition. *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 2021. 2

[34] S. Min, W. Kong, R.-C. Tu, D. Gong, C. Cai, W. Zhao, C. Liu, S. Zheng, H. Wang, Z. Li, et al. Hunyuan_tvr for text-video retrivial. *arXiv preprint arXiv:2204.03382*, 2022. 1

[35] M. Morales, S. Scherer, and R. Levitan. A linguistically-informed fusion approach for multimodal depression detection. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 13–24, 2018. 3

[36] J. Ni, X. Ma, L. Xu, and J. Wang. An image recognition method based on multiple bp neural networks fusion. In *Proceedings of the International Conference on Information Acquisition*, pages 323–326, 2004. 2

[37] Panayotov, Vassil, C. Guoguo, P. Daniel, and K. Sanjeev. Librispeech: An asr corpus based on public domain audio books. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210, 2015. 9

[38] J. Paul. The distribution of the flora of the alpine zone. pages 37–50, 1912. 6

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021. 1

[40] A. Rohrbach, M. Rohrbach, and B. Schiele. The long-short story of movie description. In *German conference on pattern recognition*, pages 209–221. Springer, 2015. 7

[41] H. Shalu, P. M. Harikrishnan, C. Harisankar, A. Das, S. Majumder, A. Datar, M. SubinMathew, A. Das, and J. Kadiwala. Depression status estimation by deep learning based hybrid multi-modal fusion model. *ArXiv*, abs/2011.14966, 2020. 3

[42] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the ACM International Conference on Multimedia*, page 399–402, 2005. 2

[43] X. Song, J. Chen, Z. Wu, and Y.-G. Jiang. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia*, 24:2914–2923, 2022. 1

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems*, page 6000–6010, 2017. 2

[45] M. Wang, G.-W. Yang, S.-M. Hu, S.-T. Yau, and A. Shamir. Write-a-video: Computational video montage from themed text. *ACM Transactions on Graphics*, 38(6):177:1–177:13, 2019. 2

[46] Q. Wang, Y. Zhang, Y. Zheng, P. Pan, and X.-S. Hua. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*, 2022. 4, 9

[47] W. Wang, J. Gao, X. Yang, and C. Xu. Learning coarse-

to-fine graph neural networks for video-text retrieval. *IEEE Transactions on Multimedia*, 23:2386–2397, 2021. 1

[48] W. Wang, J. Gao, X. Yang, and C. Xu. Many hands make light work: Transferring knowledge from auxiliary tasks for video-text retrieval. *IEEE Transactions on Multimedia*, pages 1–1, 2022. 1

[49] X. Wang, R. B. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2

[50] X. Wang, L. Zhu, and Y. Yang. T2vlad: Global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5075–5084, 2021. 9

[51] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1583–1597, 2016. 2

[52] Z. Wu, L. Cai, and H. Meng. Multi-level fusion of audio and visual features for speaker identification. In *Advances in Biometrics*, volume 3832, pages 493–499, 2006. 3

[53] S. Xie, C. Sun, J. Huang, Z. Tu, and K. P. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision*, 2018. 5

[54] H. Xu and T.-S. Chua. Fusion of av features and external information sources for event detection in team sports video. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2:44–67, 2006. 2

[55] H. Xu, W. Liu, J. Liu, M. Li, Y. Feng, Y. Peng, Y. Shi, X. Sun, and M. Wang. Hybrid multimodal fusion for humor detection. *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 2022. 3

[56] H. Xu, R. Zeng, Q. Wu, M. Tan, and C. Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the ACM International Conference on Multimedia*, pages 3893–3901, 2020. 2

[57] J. Xu, T. Mei, T. Yao, and Y. 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. 7

[58] Y. Yu, J. Kim, and G. Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 471–487, 2018. 7

[59] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1821–1830, 2017. 2

[60] B. Zhang, H. Hu, and F. Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision*, pages 374–390, 2018. 7

[61] H. Zhang, Y. Yang, F. Qi, S. Qian, and C. Xu. Robust video-text retrieval via noisy pair calibration. *IEEE Transactions on Multimedia*, pages 1–14, 2023. 1

[62] J.-Q. Zhang, X. Xu, Z.-M. Shen, Z. Huang, Y. Zhao, Y.-P. Cao, P. Wan, and M. Wang. Write-an-animation: High-level text-based animation editing with character-scene interaction. *Computer Graphics Forum*, 40, 2021. 2

[63] P. Zhang, D. Wang, and H. Lu. Multi-modal visual tracking: Review and experimental comparison. *Computational Visual Media*, 2024. 2, 5