SAVE: Speech-Aware Video Representation Learning for Video-Text Retrieval

Ruixiang Zhao* Zhihao Xu* Bangxiang Lan Zijie Xin Jingyu Liu Xirong Li†
Renmin University of China

https://github.com/ruc-aimc-lab/SAVE

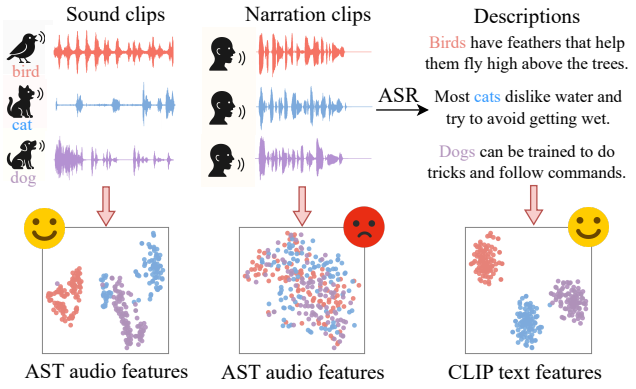
Abstract

For video-text retrieval, the use of CLIP has been a de facto *choice. Since CLIP provides only image and text encoders, this consensus has led to a biased paradigm that entirely ignores the sound track of videos. While several attempts have been made to reintroduce audio – typically by incorporating an audio encoder and fusing its output with visual features – these methods face two challenges:* ***ineffective representation of speech content*** *and* ***suboptimal vision-audio fusion****. To address these issues jointly, we propose* ***SAVE****, a* ***S****peech* ***A****ware* ***V****ideo* ***r****E****p****resentation learning method. SAVE improves upon AVIGATE, a SOTA audiovisual method, with a dedicated speech branch for more effective speech embedding. Furthermore, we introduce soft-ALBEF for early vision-audio alignment that facilitates fusion. Extensive experiments on five benchmarks show that SAVE compares favorably against the SOTA, outperforming AVIGATE by +4.1% on MSRVTT-9k, +1.9% on MSRVTT-7k, +2.5% on VATEX, +9.8% on Charades, and +2.1% on LSMDC, in light of the SumR metric.*

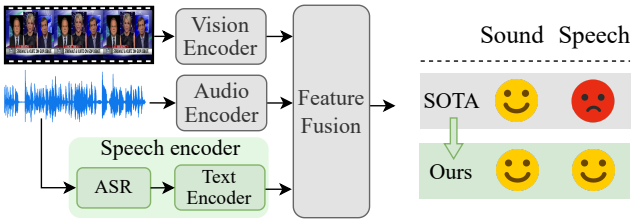
1. Introduction

Video-text retrieval (VTR), a fundamental task in vision-language understanding, aims to retrieve the most semantically relevant video w.r.t. a given text query or vice versa. Owing to the remarkable success of Contrastive Language-Image Pre-training (CLIP) [35] in bridging visual and textual modalities, the use of CLIP has been a *de facto* choice for VTR [10, 23, 32, 45, 50]. Since CLIP provides only image and text encoders, these CLIP-based methods naturally **neglect the sound track** of videos.

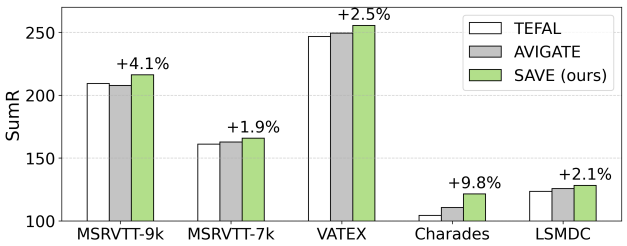
In order to reintroduce the sound track, several attempts have been made [17, 18, 27], see Tab. 1. They typically place an audio encoder alongside the image encoder for audio embedding. Subsequently, audio-enhanced video embedding is obtained by fusing the audio embedding with its visual counterpart from the vision encoder. Concerning the



(a) Limitation of current audio encoders



(b) Proposed SAVE method for speech-aware video representation learning



(c) Comparing with SOTA audio-enhanced models

Figure 1. **An overview of this paper**. **(a) Problem**: Current audio encoders (ResNet-18 [13] and AST [9]), trained on datasets of environmental sounds, are not well suited for speech embedding. **(b) Solution**: We improve the state-of-the-art in audio-enhanced video-text retrieval by introducing a dedicated speech branch for speech-aware video embedding. **(c) Results**: Our method consistently outperforms SOTA audio-enhanced models (TEFAL [17] and AVIGATE [18]) across five benchmarks.

*Equal contributions.
†Corresponding author: Xirong Li (xirong@ruc.edu.cn)

Table 1. **SOTA audiovisual methods for video-text retrieval**.

Method	Audio	Speech	Video embedding network	Early vision-audio alignment	MV-9k SumR
EclipSE[27]	ResNet-18	✗	Bi-branch	✗	197.8
TEFAL[17]	AST	✗	Bi-branch	✗	209.2
AVIGATE[18]	AST	✗	Bi-branch	✗	207.7
SAVE (ours)	AST	ASR + CLIP	Tri-branch	✓	**216.2**

choice of vision-audio fusion, text-conditioned attention is considered in TEFAL [17], whilst symmetric and asymmetric cross-attention based fusions are used in EclipSE [27] and AVIGATE [18], respectively. These audiovisual methods have demonstrated superior performance compared to their vision-only counterparts.

With a closer look at the audio encoders currently used, *i.e*. ResNet-18 [13] in EclipSE and AST [9] in TEFAL and AVIGATE, we notice that they are not well suited for speech embedding. The reason is surprisingly simple: the audio encoders were not trained for speech content understanding. To demonstrate such a deficiency, we design a toy experiment to inspect whether existing audio encoders can produce well-separated embeddings for speech content. For three animal classes, *i.e*. *bird*, *cat*, and *dog*, we construct three modality-specific datasets: (i) *Sound*: Real animal vocalizations collected from AudioSet [7], (ii) *Speech*: Narration clips[1] describing characteristics per animal, and (iii) *Text*: Transcripts obtained by performing Automatic Speech Recognition (ASR) with Whisper [36] on the *Speech* dataset. Each dataset contains 100 samples per class. Sound and speech embeddings are acquired with AST, while text embeddings are produced by the CLIP text encoder. As shown in Fig. 1a, while the sound and text embeddings are well separated class-wisely, the **speech samples** are **cluttered in** the **audio feature space**[2].

Another overlooked issue in the current audiovisual methods is the difficulty in vision-audio fusion. Consider AVIGATE for instance. The method fully relies on a cross-attention based module to directly fuse the frame features (produced by the CLIP image encoder) and the audio features (produced by AST), which are misaligned by definition. Meanwhile, lessons learned in the context of vision-language pre-training show that *align before fuse* (ALBEF) [25] helps. That said, directly applying ALBEF for the vision-audio alignment is problematic. Unlike image-caption pairs used for vision-language pre-training, vision-audio pairs often lack semantic correspondence, see Fig. 2. Consequently, enforcing **early vision-audio alignment** on the noisy pairs **with hard labels** may cause the model to **learn spurious correlations** rather than genuinely meaningful alignment.

[1]We first prompt an LLM to generate text descriptions about the characteristics per animal, and then use a Text-to-Speech model to convert these texts into narration clips.

[2]Similar cluttered results are also observed in the audio feature space of Whisper, see the supplement, though it has good ASR performance.

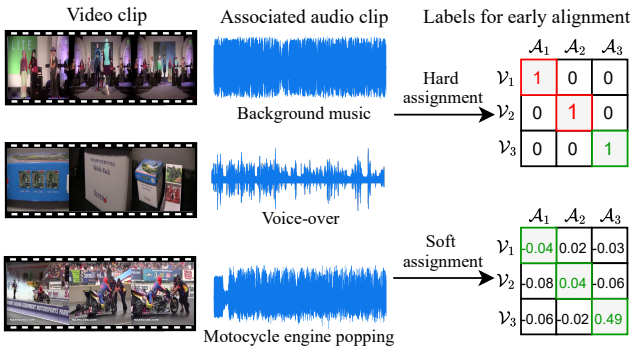


Figure 2. **Hard *vs*. soft labels for early vision-audio alignment**. For the videos in the first two rows, their associated sound tracks are not semantically relevant w.r.t. the video content. Enforcing vision-audio alignment for these videos is adverse. By contrast, the soft labels, estimated by ImageBind [8], provide finer supervision for better vision-audio alignment.

In order to jointly address the two issues, we propose in this paper a **S**peech **A**ware **V**ideo r**E**presentation learning method (**SAVE**). More specifically, for effective speech embedding, we extend the AVIGATE framework with a dedicated speech branch. Meanwhile, we develop soft-ALBEF for early vision-audio alignment, harnessing ImageBind [8] to provide soft, noise-tolerant supervision. To sum up, our main contributions are three-fold:

• **The first speech-aware video embedding for CLIP-based VTR**. We introduce a dedicated speech branch that converts narration audio into text via ASR and encodes it with a text encoder, enabling CLIP-based retrieval models to explicitly capture semantic speech information that traditional audio encoders fail to represent.

• **Soft-ALBEF for robust early vision-audio alignment**. We propose a noise-tolerant alignment strategy that leverages ImageBind to generate soft supervision signals, effectively mitigating the semantic mismatch and spurious correlations caused by directly enforcing hard alignment on noisy vision-audio pairs.

• **New state-of-the-art**. SAVE achieves superior performance across five VTR benchmarks, outperforming AVIGATE by +4.1% on MSRVTT-9k, +1.9% on MSRVTT-7k, +2.5% on VATEX, +9.8% on Charades, and +2.1% on LSMDC, in light of the SumR metric.

2. Related Work

This work aims to learn speech-aware video representation for VTR. Accordingly, we review three relevant lines of research: CLIP-based VTR, audio-enhanced VTR, and the use of speech for VTR.

CLIP-based VTR. CLIP4Clip [32] is the first to adopt CLIP for end-to-end VTR. More recent models improve the retrieval performance by introducing fine-grained interactions between the vision and text modalities, including frame-sentence alignment [4, 10, 44, 46], frame-word alignment [19, 24, 33, 37], and patch-word alignment [11, 26, 48]. Another direction of research on CLIP-based VTR emphasizes efficiency in training / inference. For efficient training, VoP [15], MV-Adapter [21] and DGL [52] resort to parameter-efficient fine-tuning. For inference efficiency, existing efforts can be roughly divided into two groups. The first group uses knowledge distillation to transfer knowledge from advanced yet computationally heavy teacher models to student models that are computationally more efficient [34, 45]. The second group is based on token compression [38, 42, 55], reducing memory footprint and computational cost by generating compact yet expressive tokens for video-text matching. The above methods do not consider the sound track.

Audio-enhanced VTR. Methods for audio-enhanced VTR put an audio encoder alongside the CLIP vision encoder [17, 18, 27]. EclipSE [27] employs ResNet-18 [13] as its audio encoder, with a symmetrical cross-attention mechanism to fuse features from the visual and audio branches. TEFAL [17] and AVIGATE [18] use Audio Spectrogram Transformer (AST) [9] as a stronger audio encoder. TEFAL adopts X-Pool [10] for both text-audio alignment and text-vision alignment. Such a dual use of X-Pool creates a large computational overhead. AVIGATE does not conduct text-audio alignment. Instead, the method integrates the audio feature into the video feature by cross-attention based gated fusion. Despite its state-of-the-art performance, AVIGATE has two drawbacks. First, current audio encoders, primarily trained for modeling environmental sounds, are ineffective in capturing speech semantics. As a consequence, the video representation of AVIGATE is not speech-aware. Second, unlike the text and vision features, which are already pre-aligned by CLIP, the audio and vision features have not undergone such pre-alignment. The lack of pre-alignment hinders the effective fusion of audio and vision features.

The Use of Speech for VTR. Before the advent of CLIP-based methods, some initial efforts were made to utilize ASR-generated text for VTR [5, 6, 29, 40, 51]. In [40, 51], text-to-video retrieval is implemented by matching a specific query with the ASR text. Later studies, such as CE [29], MMT [5], and Masked-MMT [6], simply treat the ASR text, vectorized by a pre-trained Word2Vec model, as one of the many video features to be fused. Such an ASR feature is found to be weak, practically overlooked during the feature fusion process.

Our work builds upon AVIGATE and addresses its drawbacks as follows. We add a third branch to the AVIGATE network for explicit speech content representation. Inspired by the successful use of ASR text in cross-domain product retrieval [54], we choose to encode the speech modality using CLIP's text encoder. We develop soft-ALBEF for pre-alignment of the audio and vision features before fusion. These designs lead to clear improvements over the SOTA audiovisual VTR baselines (Tab. 1).

3. Method

Given a video $\mathcal{V}$ with its soundtrack $\mathcal{A}$, we perform speech-aware video embedding via a tri-branch network. As Fig. 3 shows, the network consists of a vision branch for frame embedding, an audio branch for sound embedding, and a speech branch for spoken language embedding. For the vision and audio branches, we follow AVIGATE [18], the SOTA for learning audio-enhanced video representation. So in what follows, we first outline the AVIGATE framework (Sec. 3.1), followed by the extension of the framework to the speech branch (Sec. 3.2). We then introduce our soft-ALBEF strategy for early vision-audio embedding alignment and a joint loss for network training (Sec. 3.3).

3.1. AVIGATE in a Nutshell

We summarize the video embedding module of AVIGATE in Eq. (1). For visual token extraction, an array of m frames are uniformly sampled from the given video $\mathcal{V}$. Per frame, the visual encoder of the CLIP model [35] (CLIP$_{vis}$) is used to extract a d-dimensional frame embedding v, which is the [CLS] token of the encoder's last layer. Processing all the m frames results in a sequence of m visual tokens $\{v_i\}$.

$$\begin{cases} \{f_1, \ldots, f_m\} & \leftarrow \text{video-to-frames}(\mathcal{V}, m), \\ \{v_1, \ldots, v_m\} & \leftarrow \text{CLIP}_{vis}(\{f_1, \ldots, f_m\}), \\ \{a_1, \ldots, a_m\} & \leftarrow \text{Resampler}(\text{AST}(\mathcal{A}), m), \\ \{\hat{a}_1, \ldots, \hat{a}_m\} & \leftarrow \text{Gated-Fusion}(\{v_i\}, \{a_i\}), \\ \{\hat{v}_i\} & \leftarrow 0.95\{v_i\} + 0.05\{\hat{a}_i\}. \end{cases} \tag{1}$$

For audio token extraction, the source track $\mathcal{A}$ is first transformed into Mel filter bank features and then fed into the AST model [9], yielding a sequence of audio tokens. For token reduction, the sequence goes through a transformer-based resampler with m learnable queries, generating a shortened sequence of m d-dimensional audio tokens $\{a_i\}$. Next, a Transformer-based gated fusion is performed, with $\{v_i\}$ as Q and $\{a_i\}$ as K / V, to extract visually related audio tokens $\{\hat{a}_i\}$. A weighted sum of $\{\hat{a}_i\}$ and $\{v_i\}$ leads to $\{\hat{v}_i\}$ as audiovisual tokens. The given video is jointly represented by $\{\hat{v}_i\}$ and their mean $\bar{v}$.

For a specific text $\mathcal{T}$ to be matched with $\mathcal{V}$, its d-dimensional embedding t is obtained by the CLIP's text encoder (CLIP$_{txt}$). A global-level video-text similarity $s_g(\mathcal{V}, \mathcal{T})$ is computed as the cosine similarity between $\bar{v}$ and t. A local-level similarity $s_l(\mathcal{V}, \mathcal{T})$ is obtained by first computing the cosine similarity between $\hat{v}_i$ and t and then

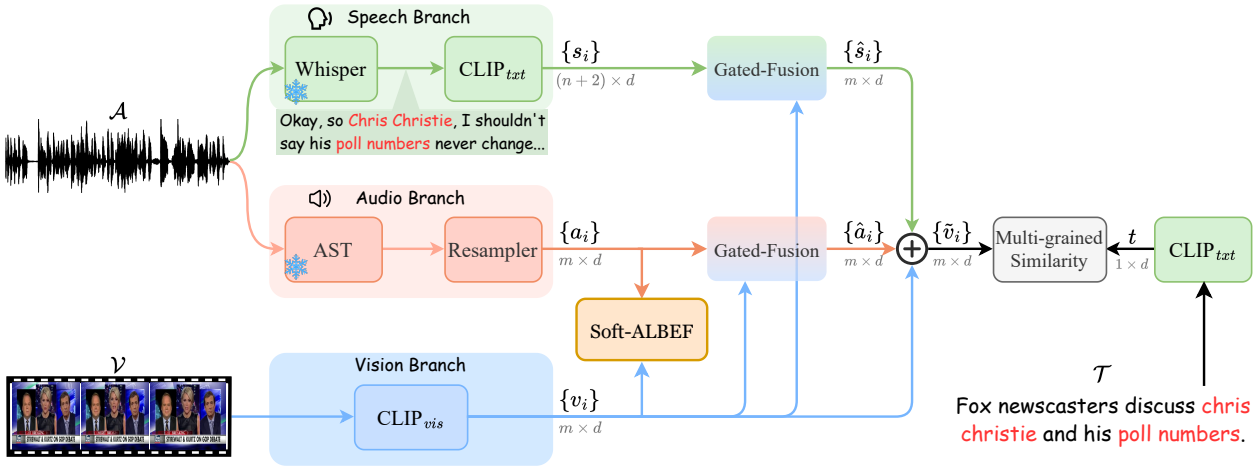


Figure 3. **Proposed speech-aware video representation learning (SAVE) method for video-text retrieval**. Given a short video $\mathcal{V}$ associated with a sound track $\mathcal{A}$, SAVE uses a tri-branch network to embed the video frames to a set of visual tokens $\{v_i\}$, the sound to a set of audio tokens $\{a_i\}$, and the speech to a set of textual tokens $\{s_i\}$. Gated fusion conditioned on the visual tokens is performed on the audio and textual tokens, yielding fused tokens $\{\hat{a}_i\}$ and $\{\hat{s}_i\}$, respectively. By aggregating $\{v_i\}$, $\{\hat{a}_i\}$ and $\{\hat{s}_i\}$, we obtain $\{\hat{v}_i\}$ as a speech-aware video representation. The video's relevance w.r.t. a specific query $\mathcal{T}$ is computed as a multi-grained similarity between $\{\hat{v}_i\}$ and the query embedding t. Soft-ALBEF is used only during training for better alignment between the visual and audio tokens.

aggregating the multiple scores using a log-sum-exp function. The final similarity is obtained as $(s_g + s_l)/2$.

3.2. Adding a Speech Branch to AVIGATE

In order to make the video representation speech-aware, we extend the AVIGATE network with a new speech branch. As shown in Fig. 3, we use Whisper large-v3 [36], a state-of-the-art ASR model, to transcribe the speech within the sound track to a sequence of n words $\{w_1, \ldots, w_n\}$. We then use CLIP$_{txt}$ for textual token extraction, obtaining a sequence of $n+2$ d-dimensional tokens as $\{s_S, s_1, \ldots, s_n, s_E\}$, where s_S and s_E are the start and end tokens automatically added by the text encoder.

We also use a Gated-Fusion module, with the visual tokens $\{v_i\}$ as Q and the speech-derived tokens as K and V, producing visually selected speech tokens $\{\hat{s}_1, \ldots, \hat{s}_m\}$. The key dataflow of our speech branch is summarized as

$$\begin{cases} \{w_1, \ldots, w_n\} & \leftarrow \text{Whisper}(\mathcal{A}), \\ \{s_S, s_1, \ldots, s_n, s_E\} & \leftarrow \text{CLIP}_{txt}(\{w_1, \ldots, w_n\}), \\ \{\hat{s}_1, \ldots, \hat{s}_m\} & \leftarrow \text{Gate-Fusion}(\{v_i\}, \{s_i\}). \end{cases} \tag{2}$$

Consequently, we obtain speech-aware audiovisual tokens $\{\tilde{v}_i\}$ as $\{v_i\} + (\{\hat{a}_i\} + \{\hat{s}_i\})/2$. Our motivation for combining the modality embeddings in this manner is three-fold. First, lacking prior knowledge of videos to be retrieved, we treat the speech and the audio branches as equally important. Second, as the visual content generally matters more, averaging the speech and audio embeddings first and then combining them with the vision embedding effectively assigns a larger weight to vision in a parameter-free manner. Third, the simple fusion encourages the Gate-Fusion modules to learn which signals truly matter.

3.3. Training with Soft-ALBEF

Soft Align before Fuse. As formalized in Eqs. (1) and (2), the Gated-Fusion modules play a crucial role in vision-audio and vision-speech feature interaction and fusion. For cross-attention based fusion to be effective, the query (Q) needs to be well aligned with the key (K). This alignment is naturally present for vision-speech fusion, where the visual tokens v_i and speech tokens s_i are pre-aligned through the shared CLIP space (CLIP$_{vis}$ and CLIP$_{txt}$). By contrast, such pre-alignment is absent for the vision-audio counterpart. Furthermore, the inherent noise in vision-sound pairs (Fig. 2) makes directly applying ALBEF [25] problematic. To address both issues, we replace ALBEF's hard labels with vision-sound relevance scores computed by ImageBind [8], a strategy we term soft-ALBEF.

Given a batch of B video-audio pairs $\{(\mathcal{V}_i, \mathcal{A}_i)\}_{i=1}^{B}$, we employ ImageBind to obtain a $B \times B$ video-audio affinity matrix M_0, with $M_0[i, j]$ indicating the cosine similarity between the ImageBind embeddings of the i-th video and j-th audio. Similarly, we compute a video-audio affinity matrix M_1 from the pre-fusion video and audio embeddings generated by the current network[3]. We use M_0 as soft supervision to guide the generation of M_1 through a Pearson

[3]The video and audio embeddings are obtained by mean pooling over $\{v\}$ and $\{a\}$, respectively.

distance loss [16], namely

$$\ell_{pearson} := \frac{1}{b}\sum_{i=1}^{b} d_p\big(\sigma(M_0[i,\cdot]), \sigma(M_1[i,\cdot])\big) + \frac{1}{b}\sum_{j=1}^{b} d_p\big(\sigma(M_0[\cdot,j]), \sigma(M_1[\cdot,j])\big) \quad (3)$$

where σ is softmax function, d_p is the Pearson distance. This choice is motivated by its invariance to separate changes in scale and location [16, 45], when compared to alternatives such as MSE or Huber loss. This property ensures that our network focuses on learning the relative ranking structure, rather than fitting absolute values, and is thus more robust and aligns better with our objective of capturing meaningful video-audio correspondences.

Loss. The Pearson distance loss serves as an auxiliary objective to the primary adaptive margin-based contrastive loss from AVIGATE. The two losses are equally combined.

Handling Missing Data. In practice, not all videos have sound tracks or ASR text available (Fig. 4). When no sound track is present, the Mel filter bank is set to zero. For cases where ASR recognition fails, an empty string is used, which is then padded as a zero vector by the tokenizer of CLIP$_{txt}$.

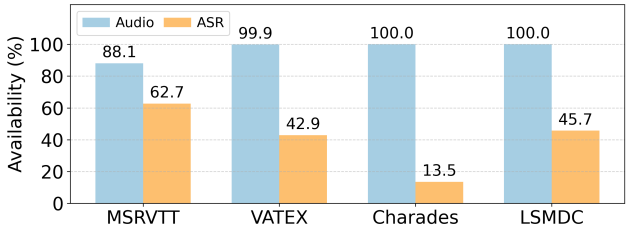


Figure 4. **Proportion of videos with audio / ASR available**.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our approach on several public datasets commonly adopted for audiovisual video-text retrieval research, including MSRVTT [49], VATEX [47], Charades [43], and LSMDC [39]. For MSRVTT, we consider both the original MSRVTT-7k split and the MSRVTT-9k partition [53]. For VATEX, we follow the partition introduced by Chen *et al*. [2]. Regarding Charades and LSMDC, we use their official data splits. A summary of dataset statistics is given in Tab. 2. Both MSRVTT-9k and MSRVTT-7k are used for the ablation study.

Evaluation criteria. We report standard rank-based retrieval metrics, *i.e*. Recall at top k ($\{1, 5, 10\}$), and SumR (R1+R5+R10) as an overall score.

Implementation details. We initialize the CLIP$_{vis}$ and CLIP$_{txt}$ with OpenAI-released CLIP [35] (ViT-B/32) weights, and AST with pretrained weights on AudioSet and ImageNet. The two CLIP$_{txt}$ encoders in speech branch and query branch share parameters during training. To reduce GPU memory overhead, the parameters of the audio encoder are kept frozen during training. The model is trained for up to 5 epochs using an Adam optimizer [22] and a cosine learning rate decay schedule [31] on 8 NVIDIA RTX 3090 GPUs. To prevent catastrophic forgetting, we finetune the CLIP backbone with a low learning rate of 1e-7, while other trainable modules are trained with a rate of 1e-4. For MSRVTT, VATEX, and LSMDC datasets, we set the max words number, max frame number and batch size to 32, 12, 128. Since videos and captions in Charades are longer and more complex, we set the max words number, max frame number and batch size to 64, 32, and 64. During testing, for datasets with validation splits (MSRVTT-7k, VATEX, LSMDC), we select the checkpoint with the best R1 on the validation set. As for MSRVTT-9k and Charades which have no validation set available, we follow [17, 18, 27, 32, 45] and report the peak test performance.

Table 2. **Number of videos per benchmark**.

Benchmark	Train	Val.	Test
MSRVTT-9k [53]	9,000	–	1,000
MSRVTT-7k [49]	6,513	497	2,990
VATEX [47]	25,991	1,500	1,500
Charades [43]	7,985	–	1,863
LSMDC [39]	101,079	7,408	1,000

4.2. Comparison with SOTA

Baselines. Both visual and audiovisual methods are compared. For the purpose of fair comparison and reproducibility of research, we only include methods that do not rely on additional training data and are open-sourced, as follows:

• *Vision-only*: CLIP4Clip [32], X-Pool [10], TS2-Net [30], X-CLIP [33], CLIP-ViP [50], STAN [28], DiCoSA [20], PromptSwitch [3], UATVR [4], UCoFiA [48], ProST [26], DGL [52], EERCF [44], TeachCLIP [45], TempMe [42], DiscoVLA [41], and PIG [23].

• *Audiovisual*: EclipSE [27], TEFAL [17], and our backbone baseline AVIGATE [18]. We further consider two AVIGATE variants: AVIGATE-h, which only uses holistic similarity s_g to generate the video-text similarity matrix, and AVIGATE+, which equips AVIGATE with our soft-ALBEF alignment strategy. We also provide a holistic variant of our SAVE, termed SAVE-h, which employs LAFF [14] to aggregate $\{\tilde{v}_i\}$ into a holistic video representation and also only computes holistic similarity.

Performance comparison. As shown in Tab. 3, AVIGATE is the state-of-the-art baseline among audiovisual methods. Our SAVE achieves new SOTA performance on text-to-video retrieval across all datasets. Specifically, SAVE outperforms AVIGATE by 8.5 on MSRVTT-9k, 3.1 on MSRVTT-7k, 6.2 on VATEX, 10.8 on Charades, and 2.6

Table 3. **Comparing with visual and audiovisual models for text-to-video retrieval**. Per baseline, we cite numbers from its original paper where applicable. The proposed SAVE model outperforms visual and audiovisual baselines on all datasets.

Model	MSRVTT-9k				MSRVTT-7k				VATEX				Charades				LSMDC				mR1
	R1	R5	R10	SumR	R1	R5	R10	SumR	R1	R5	R10	SumR	R1	R5	R10	SumR	R1	R5	R10	SumR	
Vision-only:																					
CLIP4Clip, Neucom22[32]	44.5	71.4	81.6	197.5	29.4	54.9	65.8	150.1	61.6	91.1	95.8	248.5	17.7	39.5	50.4	107.6	22.6	41.0	49.1	112.7	35.1
X-Pool, CVPR22[10]	46.9	72.8	82.2	201.9	-	-	-	-	-	-	-	-	-	-	-	-	25.2	43.7	53.5	122.4	-
TS2-Net, ECCV22[30]	47.0	74.5	83.8	205.3	29.9	56.4	67.3	153.6	59.1	90.0	95.2	244.3	16.9	38.5	49.5	104.9	23.4	42.3	50.9	116.6	35.3
X-CLIP, MM22[33]	46.1	73.0	83.1	202.2	31.2	57.4	68.1	156.7	62.2	90.9	95.4	248.5	18.9	41.3	53.0	113.2	23.3	43.0	-	-	36.3
CLIP-ViP, ICLR23[50]	46.5	72.1	82.5	201.1	30.2	56.0	69.1	155.3	62.1	90.2	97.0	249.3	-	-	-	-	-	-	-	-	-
STAN, CVPR23[28]	46.9	72.8	82.8	202.5	-	-	-	-	-	-	-	-	-	-	-	-	23.7	42.7	51.8	118.2	-
DiCoSA, IJCAI23[20]	47.5	74.7	83.8	206.0	-	-	-	-	-	-	-	-	-	-	-	-	25.4	43.6	54.0	123.0	-
PromptSwitch, ICCV23[3]	47.8	73.9	82.2	203.9	-	-	-	-	-	-	-	-	-	-	-	-	23.1	41.7	50.5	115.3	-
UATVR, ICCV23[4]	47.5	73.9	83.5	204.9	-	-	-	-	61.3	91.0	95.6	247.9	-	-	-	-	-	-	-	-	-
UCoFiA, ICCV23[48]	48.2	73.3	82.3	203.8	29.9	56.0	66.7	152.6	61.1	90.5	-	-	17.6	38.6	48.6	104.8	22.5	41.6	51.4	115.5	35.9
ProST, ICCV23[26]	48.2	74.6	83.4	206.2	30.4	56.7	67.7	154.8	60.6	90.5	95.4	246.5	17.8	37.0	48.3	103.1	24.1	42.5	51.6	118.2	36.2
DGL, AAAI24[52]	45.8	69.3	79.4	194.5	-	-	-	-	56.2	87.1	93.5	236.8	-	-	-	-	21.4	39.4	48.4	109.2	-
EERCF, AAAI24[44]	47.8	74.1	84.1	206.0	31.5	57.4	67.6	156.5	62.6	91.5	95.8	249.9	-	-	-	-	-	-	-	-	-
TeachCLIP, CVPR24[45]	46.8	74.3	82.6	203.7	30.9	57.1	68.0	156.0	63.6	91.9	96.1	251.6	19.2	40.1	51.2	110.5	-	-	-	-	-
TempMe, ICLR25[42]	46.1	71.8	80.7	198.6	-	-	-	-	-	-	-	-	-	-	-	-	23.5	41.7	51.8	117.0	-
DiscoVLA, CVPR25[41]	47.0	73.0	82.8	202.8	-	-	-	-	-	-	-	-	-	-	-	-	23.6	42.0	52.3	117.9	-
PIG, ICCV25[23]	48.6	72.8	81.6	203.0	31.8	57.3	68.0	157.1	64.0	91.5	96.6	252.1	-	-	-	-	-	-	-	-	-
Audiovisual:																					
EclipSE, ECCV22[27]	44.9	71.3	81.6	197.8	30.2	55.9	66.6	152.7	57.8	88.4	94.3	240.5	15.7	-	-	-	22.2	43.8	52.9	118.9	34.2
TEFAL, ICCV23[17]	49.4	75.9	83.9	209.2	32.6	59.1	69.3	161.0	61.0	90.4	95.3	246.7	18.5	37.3	48.6	104.4	24.7	45.1	53.7	123.5	37.2
AVIGATE, CVPR25[18]	50.2	74.3	83.2	207.7	32.7	59.8	70.2	162.7	63.1	90.7	95.5	249.3	18.8	40.0	51.8	110.6	24.6	46.0	55.1	125.7	37.9
AVIGATE-h [18]	46.1	73.4	82.4	201.9	32.0	58.6	69.1	159.7	60.8	90.1	95.4	246.3	18.6	39.5	51.5	109.6	22.8	43.7	53.5	120.0	36.1
AVIGATE+	50.9	75.8	85.2	211.9	33.0	60.3	70.8	164.1	64.8	92.3	96.6	253.7	19.9	42.3	54.4	116.6	25.5	45.7	55.1	126.3	38.8
SAVE	**51.3**	**78.0**	**86.9**	**216.2**	**33.5**	**60.9**	**71.4**	**165.8**	**66.1**	**92.6**	**96.8**	**255.5**	**20.8**	**44.7**	**55.9**	**121.4**	**26.1**	**46.4**	**55.8**	**128.3**	**39.6**
SAVE-h	49.2	77.0	85.1	211.3	32.9	60.0	70.4	163.3	64.5	91.6	96.0	252.1	19.3	41.2	53.5	114.0	24.4	43.8	53.4	121.6	38.1

on LSMDC in SumR, and 1.7 in mR1 (mean R1 across all datasets). The consistent improvements across different datasets validate the effectiveness of our method. Notably, despite only 13.5% of videos in the Charades dataset containing ASR transcripts, our SAVE still achieves a substantial 10.8 SumR improvement over AVIGATE. We attribute it to our soft-ALBEF alignment strategy that better leverages sound information in the audio modality. Comparing SAVE and AVIGATE+ reveals the gain of the speech branch. Our SAVE-h variant also yields comparable performance using only holistic video representations, further demonstrating the robustness of our design.

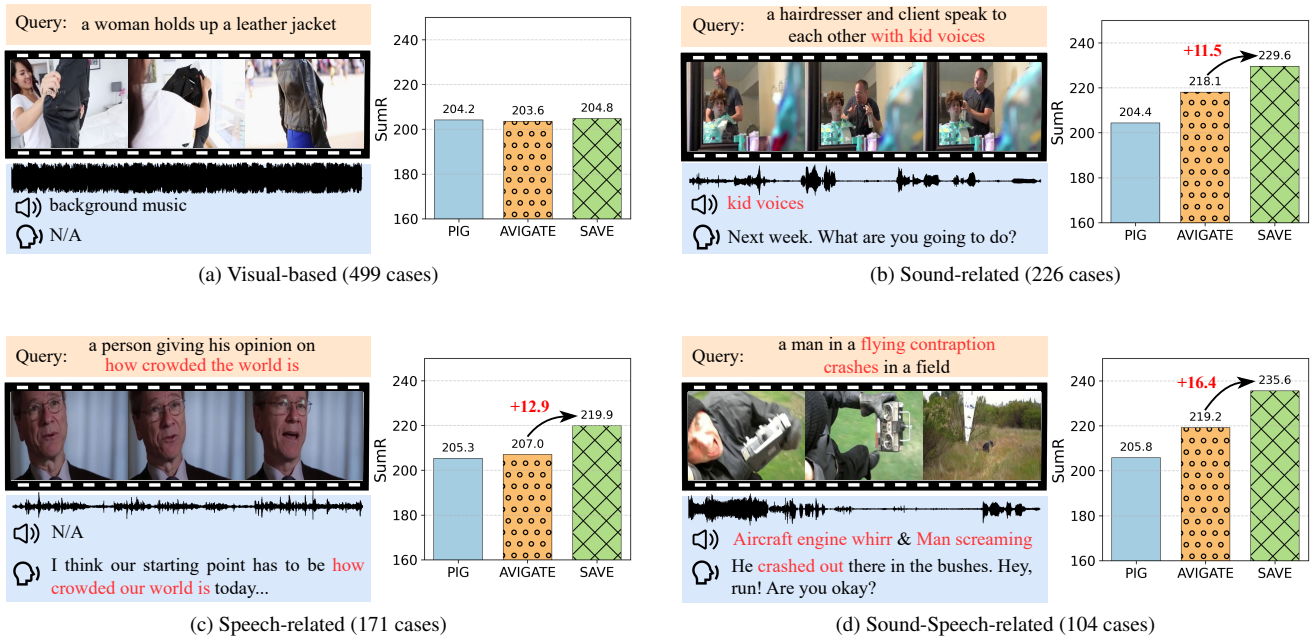
Efficiency Analysis. Let $n_{\mathcal{T}}$, $n_{\mathcal{V}}$, $n_{\mathcal{A}}$, and $n_{\mathcal{S}}$ be the number of text queries, videos, audios, and speech transcripts, respectively, where $n_{\mathcal{V}} \geq n_{\mathcal{A}} \geq n_{\mathcal{S}}$. As shown in Tab. 4, the computational complexity reflects the theoretical scaling behavior w.r.t. the number of multimodal samples, while the inference time measures the actual online latency for a single text query embedding extraction and similarity computation between the query embedding and all pre-extracted video representations. Methods such as X-Pool and TEFAL rely on query-dependent interactions, resulting in quadratic computational complexity. In contrast, SAVE maintains linear complexity, ensuring scalability to large video collections. Despite the additional speech branch, SAVE achieves the same inference latency as AVIGATE, as all video features can be extracted offline and thus do not contribute to the online inference cost.

Table 4. **Efficiency analysis**. Inference time includes both query feature extraction and video-text similarity computation. Test set: MSRVTT-9k. GPU: NVIDIA RTX 3090.

Methods	Computational complexity	Inference time (ms)↓	SumR↑
Visual:			
CLIP4Clip [32]	$\mathcal{O}(n_{\mathcal{V}} + n_{\mathcal{T}})$	9.76	197.5
X-Pool [10]	$\mathcal{O}(n_{\mathcal{V}} n_{\mathcal{T}})$	66.31	201.9
PIG [23]	$\mathcal{O}(n_{\mathcal{V}} + n_{\mathcal{T}})$	9.76	203.0
Audiovisual:			
TEFAL [17]	$\mathcal{O}(n_{\mathcal{A}} n_{\mathcal{T}} + n_{\mathcal{V}} n_{\mathcal{T}})$	140.57	209.2
AVIGATE [18]	$\mathcal{O}(n_{\mathcal{A}} + n_{\mathcal{V}} + n_{\mathcal{T}})$	9.90	207.7
AVIGATE-h [18]	$\mathcal{O}(n_{\mathcal{A}} + n_{\mathcal{V}} + n_{\mathcal{T}})$	9.76	201.9
SAVE	$\mathcal{O}(n_{\mathcal{S}} + n_{\mathcal{A}} + n_{\mathcal{V}} + n_{\mathcal{T}})$	9.90	**216.2**
SAVE-h	$\mathcal{O}(n_{\mathcal{S}} + n_{\mathcal{A}} + n_{\mathcal{V}} + n_{\mathcal{T}})$	9.76	211.3

4.3. Understanding SAVE

SAVE achieves superior performance through a dedicated speech branch and soft-ALBEF early vision-audio alignment. To better understand the two designs, we provide a series of ablation studies as follows.



(a) Visual-based (499 cases)

(b) Sound-related (226 cases)

(c) Speech-related (171 cases)

(d) Sound-Speech-related (104 cases)

Figure 5. **Comparison per group**. Our SAVE consistently outperforms PIG (best visual model) and AVIGATE (best audiovisual model) across all groups, with the largest gain obtained in the Sound-Speech-related group.

Comparison per group. We begin with a fine-grained group comparison. To that end, we manually categorize the test samples of MSRVTT-9k into four mutually exclusive groups, depending on whether a specific query is related to sound and/or speech: (a) visual-based, (b) sound-related, (c) speech-related, and (d) sound-speech-related. As shown in Fig. 5, SAVE consistently outperforms PIG and AVIGATE across all groups. For the sound-related group, SAVE achieves an 11.5 improvement in SumR over AVIGATE, primarily stemming from our proposed soft-ALBEF alignment strategy. For the speech-related group, the substantial gain of 12.9 over AVIGATE arises from the newly introduced speech branch that explicitly models semantic information in dialogue. Notably, in the sound-speech-related group where both acoustic and semantic cues jointly contribute, SAVE achieves the largest improvement of 16.4 over AVIGATE.

Speech and sound branches both benefit video-text retrieval. As shown in Tab. 5, removing the speech branch (Setup#1) and sound branch (Setup#4) results in performance drops of 4.3 and 8.7 points in SumR on MSRVTT-9k, respectively. These results demonstrate that both branches contribute meaningfully to the model's retrieval performance. While the degradation caused by removing the sound branch appears larger, this is mainly because MSRVTT-9k contains notably more sound-related queries than speech-related ones (see Fig. 5).

Choice of ASR models. We replace Whisper with an alternative ASR model, SenseVoice [1], to evaluate the impact of ASR backbone choice. Setup#2 shows that Whisper is the optimal choice for our framework.

Table 5. **Ablation studies of SAVE**.

#	Setup	MSRVTT-9k		MSRVTT-7k	
		R1	SumR	R1	SumR
0	Full	**51.3**	**216.2**	**33.5**	**165.8**
Speech branch:					
1	*w/o* Speech Branch	50.9	211.9(4.3↓)	33.0	164.1(1.7↓)
2	ASR: Whisper → SenseVoice	51.2	215.4(0.8↓)	33.5	165.4(0.4↓)
3	CLIP_{txt}: *w/o* parameter sharing	51.0	214.8(1.4↓)	33.4	165.0(0.8↓)
Sound branch:					
4	*w/o* Sound Branch	48.5	207.5(8.7↓)	32.0	159.3(6.5↓)
5	*w/o* Soft-ALBEF	49.3	213.8(2.4↓)	33.3	164.8(1.0↓)
6	Soft-ALBEF → ALBEF	49.2	211.2(5.0↓)	33.2	164.6(1.2↓)
7	Pearson → MSE	50.8	215.0(1.2↓)	33.4	165.3(0.5↓)
8	Pearson → Huber	50.8	215.7(0.5↓)	33.5	165.5(0.3↓)
9	ImageBind → AudioCLIP	51.1	215.8(0.4↓)	33.6	165.6(0.2↓)
Fusion:					
10	Late-fusion	47.8	206.5(9.7↓)	31.4	155.6(10.2↓)
11	Learnable weights	51.2	215.5(0.7↓)	33.4	165.3(0.5↓)

Using the same text encoder for query and speech embedding? Since the text encoder in the speech branch and that in the query encoder are both instantiated by CLIP, a natural question is whether they shall share the same weights. The results of Setup#3 indicates that sharing is preferred. Our interpretation is that parameter sharing helps reduce the gap between the speech and query modalities,

thereby facilitating more effective cross-modal retrieval.

The role of early vision-audio alignment. Directly applying ABLEF for vision-audio alignment even leads to performance degradation compared with no alignment (Setup #5 *vs* #6). This indicates that making video-audio pairs for early alignment is non-trivial. By contrast, our soft-ABLEF strategy offers a feasible solution.

Comparison with data filtering. A simpler alternative to soft-ALBEF is to filter out *low-ImageBind-score* video-audio pairs before applying ALBEF. As shown in Fig. 6, soft-ALBEF consistently outperforms ALBEF across all filtered-out ratios, demonstrating its effectiveness.

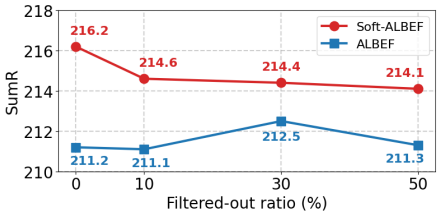


Figure 6. Soft-ALBEF *vs* ALBEF with *low-score* data filtered out.

Which loss for soft-ALBEF? As shown in Tab. 5, the results of Setup#7 and Setup#8 indicate that Pearson distance loss is a better choice than MSE and Huber loss.

Choice of teacher model. Setup#9 replaces ImageBind with a weaker teacher AudioCLIP [12], and still yields a clear gain over AVIGATE, confirming that the benefit comes from soft-labeling rather than the specific teacher.

Late fusion *vs*. early fusion. Setup#10 implements a late-fusion baseline, where the speech-query similarity score is computed separately and averaged with the video-audio-query similarity score. The performance drop demonstrates the superiority of our early fusion strategy.

Learnable weights for obtaining $\{\tilde{v}_i\}$. Setup#11 replaces the parameter-free fusion $\{v_i\}+(\{\hat{a}_i\}+\{\hat{s}_i\})/2$ with a parameterized version $\alpha\{v_i\}+\beta\{\hat{a}_i\}+(1-\alpha-\beta)\{\hat{s}_i\}$ using learnable weights α and β. The performance does not improve, validating our motivation discussed in Sec. 3.2.

Using ImageBind directly for VTR? We employ ImageBind to generate soft labels for soft-ALBEF. A natural question arises as *Is ImageBind itself suitable for VTR*? We evaluate ImageBind under three different modality configurations for constructing video representations:

• ImageBind-v: Video representation is derived from the visual modality by averaging frame-level features.

• ImageBind-va: Video representation is obtained by combining visual and audio features.

• ImageBind-vas: Incorporates video, audio, and speech features jointly for video representation.

Note that ImageBind uses ViT-H/14, which is too large to be fully finetuned for VTR. To investigate its applicability within our GPU budget (8× RTX 3090 GPUs). We evaluate all three variants in a zero-shot manner. Additionally, for ImageBind-vas, we further explore a partial finetuning strategy, where the vision and audio encoders are frozen except for their linear projection layers, while the text encoder, computationally lightweight, is fully finetuned.

As shown in Tab. 6, ImageBind performs worse than our SAVE (ViT-B/32) under both settings, suggesting that ImageBind is not directly suitable as a backbone for video-text retrieval, especially under limited computational resources. However, this conclusion does not conflict with its specialized use in our soft-ALBEF strategy. For soft-ALBEF, we do not use ImageBind as the primary retrieval backbone. Instead, we leverage its powerful vision-audio alignment capabilities only to compute a high-quality similarity matrix.

Table 6. **Comparison with ImageBind**. Dataset: MSRVTT-9k.

Methods	R1	R5	R10	SumR
Zero-shot:				
ImageBind-v	39.1	62.8	73.1	175.0
ImageBind-va	39.4	63.2	73.7	176.3
ImageBind-vas	40.6	62.6	74.1	177.3
Finetuned:				
ImageBind-vas	47.2	72.9	82.0	202.1
SAVE	**51.3**	**78.0**	**86.9**	**216.2**

5. Conclusions

We propose SAVE, a speech-aware video representation for audio-enhanced video-text retrieval (VTR). Our work identifies two key limitations in current methods for audio-enhanced VTR, namely the under-exploitation of speech semantics and the absence of early video-audio alignment, and accordingly proposes two targeted, effective solutions. Extensive experiments on multiple benchmarks allow us to conclude as follows. The dedicated speech branch effectively captures semantic information from dialog and provides complementary benefits to the sound branch. The soft-ALBEF strategy offers an effective solution for early vision-audio alignment when direct pairing is noisy or unreliable. This study opens a promising direction for the integration of speech with other types of audio cues for multimodal VTR.

Limitation of this study. Our experiments are conducted on short video clips with brief ASR transcripts that can be directly handled by the CLIP text encoder. In more complex scenarios, such as e-commerce livestream understanding, ASR transcripts are typically much longer and noisier. How to efficiently extract key information from such lengthy and noisy speech to obtain robust speech-aware video representation warrants further research.

Acknowledgments. This research was supported by NSFC (No. 62576348) and Beijing Natural Science Foundation (No. L254039).

References

[1] Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al. FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs. *arXiv*, 2024. 7

[2] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, 2020. 5

[3] Chaorui Deng, Qi Chen, Pengda Qin, Da Chen, and Qi Wu. Prompt Switch: Efficient CLIP adaptation for text-video retrieval. In *ICCV*, 2023. 5, 6, 3

[4] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. UATVR: Uncertainty-adaptive text-video retrieval. In *ICCV*, 2023. 3, 5, 6

[5] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 3

[6] Valentin Gabeur, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid. Masking modalities for cross-modal video retrieval. In *WACV*, 2022. 3

[7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 2

[8] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *CVPR*, 2023. 2, 4, 1

[9] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Interspeech*, 2021. 1, 2, 3

[10] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-Pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, 2022. 1, 3, 5, 6

[11] Peiyan Guan, Renjing Pei, Bin Shao, Jianzhuang Liu, Weimian Li, Jiaxi Gu, Hang Xu, Songcen Xu, Youliang Yan, and Edmund Y Lam. PIDRo: Parallel isomeric attention with dynamic routing for text-video retrieval. In *ICCV*, 2023. 3

[12] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *ICASSP*, 2022. 8

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2, 3

[14] Fan Hu, Aozhu Chen, Ziyue Wang, Fangming Zhou, Jianfeng Dong, and Xirong Li. Lightweight attentional feature fusion: A new baseline for text-to-video retrieval. In *ECCV*, 2022. 5

[15] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. VoP: Text-video co-operative prompt tuning for cross-modal retrieval. In *CVPR*, 2023. 3

[16] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *NeurIPS*, 2022. 5

[17] Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mohamed Omar. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *ICCV*, 2023. 1, 2, 3, 5, 6

[18] Boseung Jeong, Jicheol Park, Sungyeon Kim, and Suha Kwak. Learning audio-guided video representation with gated attention for video-text retrieval. In *CVPR*, 2025. 1, 2, 3, 5, 6

[19] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *CVPR*, 2023. 3

[20] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. Text-video retrieval with disentangled conceptualization and set-to-set alignment. In *IJCAI*, 2023. 5, 6, 3

[21] Xiaojie Jin, Bowen Zhang, Weibo Gong, Kai Xu, Xueqing Deng, Peng Wang, Zhao Zhang, Xiaohui Shen, and Jiashi Feng. MV-Adapter: Multimodal video transfer learning for video text retrieval. In *CVPR*, 2024. 3

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 5

[23] Bangxiang Lan, Ruobing Xie, Ruixiang Zhao, Xingwu Sun, Zhanhui Kang, Gang Yang, and Xirong Li. Hybrid-Tower: Fine-grained pseudo-query interaction and generation for text-to-video retrieval. In *ICCV*, 2025. 1, 5, 6, 3

[24] Huy Le, Nhat Chung, Tung Kieu, Anh Nguyen, and Ngan Le. BiMa: Towards biases mitigation for text-video retrieval via scene element guidance. In *ACMMM*, 2025. 3

[25] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align Before Fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021. 2, 4

[26] Pandeng Li, Chen-Wei Xie, Liming Zhao, Hongtao Xie, Jiannan Ge, Yun Zheng, Deli Zhao, and Yongdong Zhang. Progressive spatio-temporal prototype matching for text-video retrieval. In *ICCV*, 2023. 3, 5, 6

[27] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. EclipSE: Efficient long-range video retrieval using sight and sound. In *ECCV*, 2022. 1, 2, 3, 5, 6

[28] Ruyang Liu, Jingjia Huang, Ge Li, Jiashi Feng, Xinglong Wu, and Thomas H Li. Revisiting temporal modeling for CLIP-based image-to-video knowledge transferring. In *CVPR*, 2023. 5, 6, 3

[29] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 3

[30] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. TS2-Net: Token shift and selection transformer for text-video retrieval. In *ECCV*, 2022. 5, 6, 3

[31] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 5

[32] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 2022. 1, 3, 5, 6

[33] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACMMM*, 2022. 3, 5, 6

[34] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. CLIPPING: Distilling CLIP-based models with a student base for video-language retrieval. In *CVPR*, 2023. 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 3, 5

[36] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023. 2, 4, 1

[37] Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M de Melo, Benjamin Van Durme, and Rama Chellappa. Video-ColBERT: Contextualized late interaction for text-to-video retrieval. In *CVPR*, 2025. 3

[38] Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou. TESTA: Temporal-spatial token aggregation for long-form video-language understanding. In *EMNLP*, 2023. 3

[39] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015. 5

[40] Ken Schmidt, Thomas Körner, Stephan Heinich, and Thomas Wilhelm. A two-step approach to video retrieval based on ASR transcriptions. *MediaEval*, 2011. 3

[41] Leqi Shen, Guoqiang Gong, Tianxiang Hao, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, Jungong Han, and Guiguang Ding. DiscoVLA: Discrepancy reduction in vision, language, and alignment for parameter-efficient video-text retrieval. In *CVPR*, 2025. 5, 6, 3

[42] Leqi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. TempMe: Video temporal token merging for efficient text-video retrieval. In *ICLR*, 2025. 3, 5, 6

[43] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 5

[44] Kaibin Tian, Yanhua Cheng, Yi Liu, Xinglin Hou, Quan Chen, and Han Li. Towards efficient and effective text-to-video retrieval with coarse-to-fine visual representation learning. In *AAAI*, 2024. 3, 5, 6

[45] Kaibin Tian, Ruixiang Zhao, Zijie Xin, Bangxiang Lan, and Xirong Li. Holistic features are almost sufficient for text-to-video retrieval. In *CVPR*, 2024. 1, 3, 5, 6

[46] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval. *arXiv*, 2022. 3

[47] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 5

[48] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *ICCV*, 2023. 3, 5, 6

[49] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016. 5

[50] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. CLIP-ViP: Adapting pre-trained image-text model to video-language alignment. In *ICLR*, 2023. 1, 5, 6

[51] Haojin Yang and Christoph Meinel. Content based lecture video retrieval using speech and video text information. *TLT*, 2014. 3

[52] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. DGL: Dynamic global-local prompt tuning for text-video retrieval. In *AAAI*, 2024. 3, 5, 6

[53] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 5

[54] Ruixiang Zhao, Jian Jia, Yan Li, Xuehan Bai, Quan Chen, Han Li, Peng Jiang, and Xirong Li. ASR-enhanced multimodal representation learning for cross-domain product retrieval. *TMM*, 2026. 3

[55] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. CenterCLIP: Token clustering for efficient text-video retrieval. In *SIGIR*, 2022. 3