Text-to-Multimodal Retrieval with Bimodal Input Fusion in Shared Cross-Modal Transformer

Pranav Arora, Selen Pehlivan, Jorma Laaksonen
Aalto University, VTT Technical Research Centre of Finland, Aalto University
Espoo Finland, Oulu Finland, Espoo Finland
pranav.arora@aalto.fi, selen.pehlivantort@vtt.fi, jorma.laaksonen@aalto.fi

Abstract

The rapid proliferation of multimedia content has necessitated the development of effective multimodal video retrieval systems. Multimodal video retrieval is a non-trivial task involving retrieval of relevant information across different modalities, such as text, audio, and visual. This work aims to improve multimodal retrieval by guiding the creation of a shared embedding space with task-specific contrastive loss functions. An important aspect of our work is to propose a model that learns retrieval cues for the textual query from multiple modalities both separately and jointly within a hierarchical architecture that can be flexibly extended and fine-tuned for any number of modalities. To this end, the loss functions and the architectural design of the model are developed with a strong focus on increasing the mutual information between the textual and cross-modal representations. The proposed approach is quantitatively evaluated on the MSR-VTT and YouCook2 text-to-video retrieval benchmark datasets. The results showcase that the approach not only holds its own against state-of-the-art methods, but also outperforms them in a number of scenarios, with a notable relative improvements from baseline in R@1, R@5 and R@10 metrics.

Keywords: text-to-video retrieval, multimodal retrieval, modality fusion, transfer learning, contrastive learning, multimodal transformers, cross-modality

1. Introduction

The exponential growth of multimedia content has accentuated the need for robust text-to-video retrieval systems with practical applications including web search engines and personal media indexing (Lew et al., 2006; Hu et al., 2011; Liu et al., 2021; Zhu et al., 2023; Qiu, 2022). Specifically, the main user expectation from these systems, which seek to locate specific video files based on text queries, is to enhance the user experience by delivering more relevant search results and addressing also queries where mere textual information might be lacking. With the presence of various input modalities, solely relying on cross-modal understanding between text and vision (Zhao et al., 2022) is sub-optimal. Leveraging multimodal retrieval ensures richer content representation, adept handling of ambiguity, and a contextual understanding of user queries. This results in a more precise and contextually accurate media retrieval, fulfilling nuanced user expectations in the vast multimedia landscape.

Particularly, audio, being the most prevalent third modality alongside textual and visual, warrants further exploration in cross-modality applications (Zhao et al., 2022; Shvetsova et al., 2022; Chen et al., 2023a). For instance, as illustrated in Figure 1, the absence of clear visual cues for the "*...inviting his colleagues to join him*" part of the textual description can hinder the accurate learning of the required text-to-video associations. However, a trimodal embedding that also incorporates the "*invitation*" in the audio modality can enhance retrieval outcomes. Similarly, a user searching for "*romantic scene with a sunset and a soft song*" expects the system to recognize both the visual beauty of the sunset and the auditory cue of a soft song to retrieve the right video. In such cases, multimodal learning that integrates text, audio, and visual becomes indispensable for accurate retrieval.

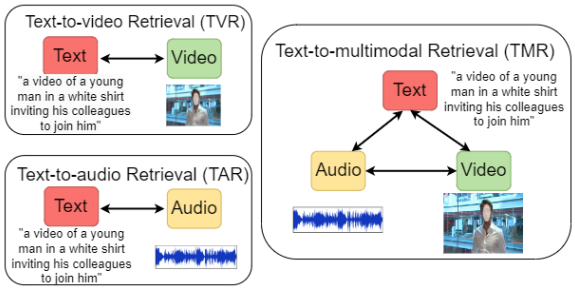

Figure 1: An example of a textual query and a video target that can be retrieved with multimodal but not with unimodal associations.

Fusion techniques are essential to multimodal learning, ensuring effective integration of information across modalities for a unified model (Chen et al., 2020; Bao et al., 2022). These techniques are categorized by Nagrani et al. (2021) as *early*, *mid*, and *late fusion*. While *early fusion* exchanges cross-modal information at the outset, *late fusion* waits until after the classifier for exchange. Recognizing the potential of *mid fusion*, the hierarchical fusion approach we propose in this paper initially processes each modality with its dedicated encoder.

LREC-COLING 2024, pages 15823–15834
20-25 May, 2024.

This is then followed by a multimodal fusion transformer that reveals shared embedding representations among the modality pairs. This process mirrors the human tendency to process individual information before integrating it, thereby enhancing text-to-multimodal retrieval (TMR) efficacy.

Another important factor for effective cross-modal integration is how transformers for multimodal fusion are designed. It has been shown that dual-encoders behave superior to fusion-encoders (Lu et al., 2019) for bimodal retrieval tasks. In these setups, modality interactions are jointly encoded via similarity scores for retrieval tasks, e.g. image-to-text retrieval (Radford et al., 2021; Li et al., 2021; Bao et al., 2022). Although bimodal architectures are well studied, an increasing number of modalities comes with design challenges. For instance, the multimodal fusion transformer for trimodality in (Shvetsova et al., 2022) causes exponential growth with the increasing number of modalities, unless additional training strategies, such as random dropping in modality combinations, are used. As Figure 2 demonstrates, we formulate in this paper a cross-attention mechanism within a scalable architecture to encode shared embeddings among all modality pairs. We believe that this is further extendable to even more modalities without changing the multimodal fusion transformer.

Our proposed approach employs cross-attention to process multiple modality pairs simultaneously, facilitating effective multimodal learning. In the context of trimodality, merging audio and visual modalities presents challenges. However, when these modalities are meticulously aligned in a shared space, they can together yield powerful retrieval cues. In summary, our contributions are:

1. For multimodal retrieval, we present a hierarchical architecture[1] that initially cultivates modal-specific unimodal experts. This is then complemented by a dedicated cross-attention fusion transformer to establish a modal-agnostic multimodal space.

2. We highlight the potential of fine-tuning loss variations to boost performance in text query based multimodal retrieval tasks.

3. We demonstrate the efficacy of audio-video fusion in enhancing text-based retrieval.

4. We assess the impact of text query length on the efficacy of retrieval systems.

2. Related Work

In the realm of deep learning, the evolution of multimodal retrieval research can be categorized in three main areas: vision-language (Bain et al., 2021; Arnab et al., 2021), vision-audio (Rouditchenko et al., 2020; Chung et al., 2019), and multimodal learning (Miech et al., 2019; Shvetsova et al., 2022; Chen et al., 2021). The development strides in the transformer architecture (Vaswani et al., 2017) and contrastive learning (Oord et al., 2018) have pushed the research. Transformers, having shown prowess in unimodal NLP, vision, and audio tasks like (Kenton and Toutanova, 2019; Arnab et al., 2021), naturally led to their adoption in multimodal learning as evidenced by numerous studies (Radford et al., 2021; Li et al., 2021; Bao et al., 2022). This trajectory inspired our exploration of transformer-based architectures for enhanced multimodal retrieval.

Pretrained backbones, like CLIP (Radford et al., 2021), excel in feature extraction and create a unified text and visual space, proving effective in multimodal research (Xue et al., 2022; Nagrani et al., 2022; Shvetsova et al., 2022; Bain et al., 2021). For video tasks, architectures like CLIP4CLIP (Luo et al., 2022) showcase their adaptability in the field.

A popular approach in vision-language and vision-audio is bimodal learning, where two transformer-based models are jointly pretrained to create a shared space (Bain et al., 2021; Rouditchenko et al., 2020), bridging the modalities. Besides fine-tuning, utilizing frozen pretrained models is also gaining traction. This approach not only saves computational resources, but also leverages the rich feature representations learned by the base models. Works such as (Luo et al., 2020; Xue et al., 2022) extend the transformer-based pretrained backbones, such as CLIP, trained on image-text pairs using transformers for text-to-video retrieval. However, while various methodologies for bimodal learning are being explored, video data contains a wealth of cues within its audio modality, which must be harnessed to develop robust retrieval systems.

Pioneering the domain of multimodal learning, (Aytar et al., 2016) introduced an innovative architecture trained on image-text and image-audio pairs. This marked one of the earliest endeavors to seamlessly integrate text and audio modalities in multimodal learning. In the context of multimodal retrieval tasks, recent works, such as (Chen et al., 2021; Akbari et al., 2021; Shvetsova et al., 2022; Chen et al., 2023a), have delved into diverse approaches to derive effective multimodal representations using text, visual, and audio modalities. A specific method (Radford et al., 2021) employs embeddings or *tokens* from unfrozen pretrained backbone networks to compute a *similarity matrix*. However, it's more common for the obtained input tokens to be processed using transformer-based models, like (Nagrani et al., 2022), to enhance performance.

[1] https://github.com/Pranav260/TMR

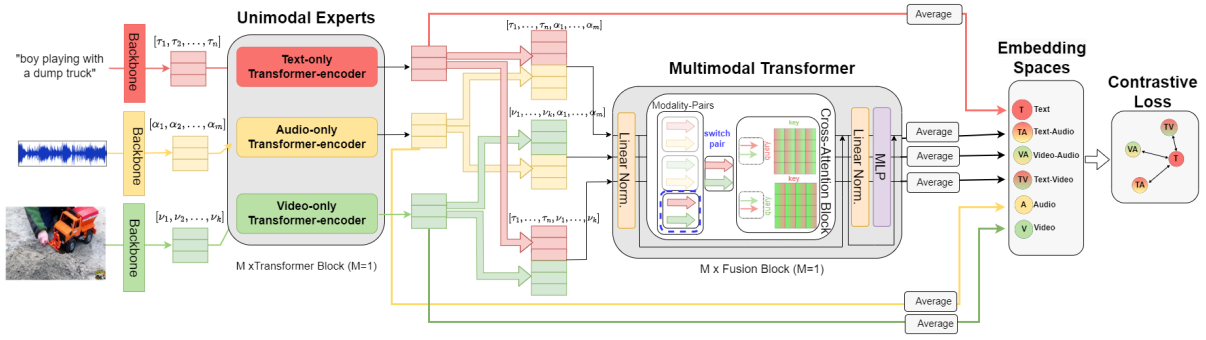

Figure 2: Our proposed model's end-to-end training pipeline for the text-to-multimodal retrieval (TMR) task. Our model uses three separate transformer-encoders as unimodal experts and a multimodal fusion transformer with shared parameters among modality pairs. The objective function for TMR is based on contrastive loss components between text t and cross-modal representations, tv, va, ta, respectively.

Delving deeper into transformer-based techniques, (Akbari et al., 2021) proposed a trimodal structure encompassing text, audio, and video. This model operates on a modality-agnostic, single-backbone Transformer by sharing weights across the three modalities. The authors of (Li et al., 2021; Chen et al., 2023a) emphasize the importance of alignment between unimodal representations to establish a robust multimodal space. Specifically, (Chen et al., 2023a) acknowledges the impact of alignment and adapts its use for various tasks, like retrieval and captioning, by adjusting the loss function to suit each task's requirements. A recent study (Ibrahimi et al., 2023) demonstrates promising performance using text-conditioned audio and visual features, without the need for pretraining on large-scale datasets.

Closest to our work, (Shvetsova et al., 2022) emphasizes the importance of immediate interaction and integration of modalities, ensuring that the cross-modal model captures the intricate interplay between them. However, it might not be able to capture the nuanced representation of single and fused representation, which offers an advantage as described in (Nagrani et al., 2021). Further using multiple contrastive losses between different modality pairs in a combinatorial manner can help in the pretraining stage but for certain tasks, specific fine-tuning strategies such as picking different combinations of losses can serve very well. These observed drawbacks serve as a foundation for building our architecture.

3. Model Architecture

Our trimodal retrieval model draws its inspiration from multimodal fusion transformers presented in recent studies such as (Xue et al., 2022; Shvetsova et al., 2022; Nagrani et al., 2022). The procedure initiates with quite standard, state-of-the-art modality-specific token generation processes. A significant enhancement in the proposed architecture is then the integration of separate transformers for each modality as unimodal experts. The unimodal experts are responsible for generating robust representations for text, audio, and video, respectively, complemented by the explicit cross-attention computation central to the multimodal fusion transformer, as depicted in Figure 2.

Our model's hierarchical structure facilitates learning retrieval cues from multiple modalities in a flexible and extensible manner. Flexibility is inherent in our approach, as we do not need to alter the existing setup to add new modalities such as depth data or haptics – simply adding another unimodal transformer suffices. We can then select the best combination to minimize contrastive losses between modalities for the desired task.

3.1. Token and Feature Extraction

Following the common practice, the inputs of all data modalities, i.e., text, audio, and video, are initially processed through backbone networks to extract features as illustrated in Figure 2. The CLIP backbone (Radford et al., 2021) is used for text and video and a trainable CNN backbone (Shvetsova et al., 2022; Rouditchenko et al., 2020) is used for audio. Subsequently, the extracted features undergo modality-specific linear projections through the token projection layers. Following the initial processing, each modality's tokens undergoes layer normalization (LN), essentially L2 normalization. The output of each LN is a three-dimensional tensor of tokens, $[B, N, C]$, where B is the batch size, N is the number of tokens, and C is the channel dimensionality. Tokens for text, audio, and video are represented as $[\tau_1, \tau_2, \ldots, \tau_n]$, $[\alpha_1, \alpha_2, \ldots, \alpha_m]$, and $[\nu_1, \nu_2, \ldots, \nu_k]$. Due to the variability in token dimensionalities across modalities, especially from the differing lengths of video clips, token normalization is employed to make the number of tokens in

each video constant, facilitating batch processing during training. To ensure consistency with prior research (Shvetsova et al., 2022; Miech et al., 2019), the same configuration in terms of input size, learning rate, and batch size is used.

3.2. Unimodal Transformers

The initial part of the model consists of three unimodal vanilla transformer-encoders for text, video, and audio as shown in Figure 2. The main aim behind using separate encoders is to generate a better representation and attune the embedding to each particular modality effectively. This enables better performance across unimodal tasks meanwhile also serving as input to the fusion transformer where these high-dimensional inputs can be really exploited in modelling cross-modal relationships.

Given tokens for text, video, and audio, each is processed by designated modality-specific unimodal transformers. The transformed representations are subsequently paired and organized as ta, tv, and va. These combinations are sequentially channeled into a fusion transformer. The outputs from the three merged modalities, in conjunction with individual outputs t, a, and v, undergo token-wise averaging and are projected into a shared embedding space.

3.3. Multimodal Fusion Transformer

The most important design decision is to craft an effective strategy for integrating any two modalities. In achieving such a fused representation, we employ a transformer-encoder that shares weights across all modality pairs. Noteworthy in this approach, and deviating from (Shvetsova et al., 2022), is the incorporation of a distinct cross-attention block as a multimodal fusion transformer. This block allows for bidirectional attention computations leveraging two cross-attention units. In this setup, a query from one modality can interact with the keys and values of another modality, and the other way around, e.g. text-to-video and video-to-text. Given the three unimodal t, a, v inputs, all bimodal inputs, i.e., ta, tv, and va, use the same cross-attention block. Our model thus has one cross-attention block shared by all modality pairs to prevent exponential growth in the number of model parameters as the number of modalities increases.

The proposed bidirectional interaction via our cross-attention block is visually represented in Figure 3. Its input and output are the stacked representation of two particular modalities, e.g. here the red and yellow blocks of tokens denoting the text and audio modalities t and a, respectively. While one cross-attention computes similarities using queries of t with keys of a, the other cross-attention computes similarities in opposite direction.

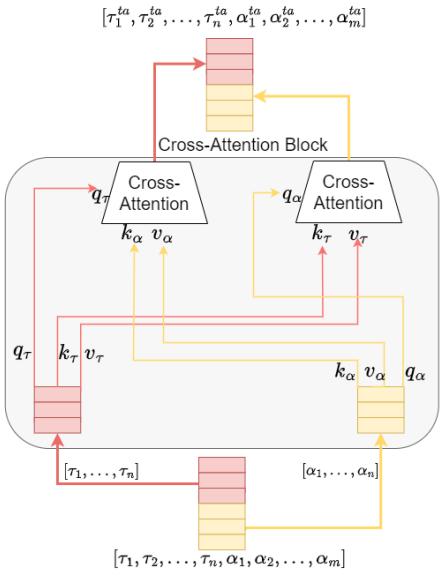

Figure 3: Detailed illustration of the cross-attention block as one layer of the multimodal fusion transformer (see Figure 2). For bimodal input, the cross-attention block jointly extracts bidirectional embeddings across modalities.

3.4. Fusion in Shared Embedding Space

Ultimately, given dedicated unimodal transformers for each modality, our model generates representations t for text, v for video, and a for audio. In contrast, the multimodal fusion transformer produces three combined representations: ta for text-audio, tv for text-video, and va for video-audio. Upon deriving the six representations, each one undergoes averaging across the token dimension N resulting in averaged dimensionality of $[B, C]$. The stacked representations are de-stacked and then individually averaged after which they are individually projected in a higher dimensionality using a modality-specific projection layer. Mathematically this is expressed for the text-video tv pair as:

$$[\tau_{i1}^{tv}, \ldots, \tau_{in}^{tv}, \nu_{i1}^{tv}, \ldots, \nu_{ik}^{tv}], \tag{1}$$

where

$$\tau_i^{tv} = \frac{1}{n}\sum_{j=1}^{n} \tau_{ij}^{tv}, \ \nu_i^{tv} = \frac{1}{k}\sum_{j=1}^{k} \nu_{ij}^{tv}. \tag{2}$$

Using the above averaged representations, these modalities are finally normalized and projected into a shared space. The projections are then element-wise added, resulting in:

$$f(\tau_i^{tv}, \nu_i^{tv}) = \sigma_t(\tau_i^{tv}) + \sigma_v(\nu_i^{tv}), \tag{3}$$

where σ_t and σ_v are the projection operations. Here, projection operations are linear layers that transform the input embedding into a shared higher-dimensional space across all modalities. Similarly

the other two stacked representations, i.e., ta and va, are converted into fused representations.

Altogether, these constitute six distinct representations, t, v, a tv, ta and va that are subsequently used for defining the contrastive loss pairs for the training.

3.5. Loss Function

As the main objective is to improve the inter-modal representations, one requires a loss function which is able to guide the model to find correlations between the modalities. In this work, we prioritize the enhancement of mutual information between textual and cross-modal representations. As a solution, we formulate the loss upon the concept of contrastive loss (Oord et al., 2018), which encourages the learning of discriminative representations by leveraging both positive and negative pairs.

Given two representations X and Y, where the similarity function used is cosine similarity between X and Y, the contrastive loss $\mathcal{L}_{XY}$ can be computed bidirectionally using the Noise Contrastive Estimation (NCE) (Oord et al., 2018) with temperature ζ and batch size B as:

$$\begin{aligned}\mathcal{L}_{XY} = &-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(X_i,Y_i)/\zeta)}{\sum_{j=1}^{B}\exp(\text{sim}(X_i,Y_j)/\zeta)} \\ &-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(X_i,Y_i)/\zeta)}{\sum_{j=1}^{B}\exp(\text{sim}(X_j,Y_i)/\zeta)}.\end{aligned} \tag{4}$$

Following (Shvetsova et al., 2022), a combination of multiple contrastive loss functions between different modalities is formulated to guide the model to convergence. However, we here focus on the TMR task and the final objective can be represented as:

$$\mathcal{L} = \sum_{\substack{X\subset\mathcal{M}_q,\\ Y\subset\mathcal{M}_r,\\ \mathcal{M}_q\cap\mathcal{M}_r=\phi}} \lambda_{XY}\mathcal{L}_{XY}, \tag{5}$$

where λ_{XY} is a weight coefficient for each loss component $\mathcal{L}_{XY}$ between two modalities. It is observed that $\mathcal{M}_r = \{t\}$ and $\mathcal{M}_q = \{ta, tv, va\}$ are powerful among various loss alternatives for the TMR task. We will study and discuss the choices and implications of these observations in detail (see Figure 4 for more insights).

4. Experimental Evaluation

4.1. Datasets and Statistics

In the pursuit of advancing multimodal learning, the choice of dataset plays a pivotal role. Our model has used the pretained weights by (Shvetsova et al., 2022) over HowTo100M dataset, and the YouCook2 and the MSR-VTT for text-to-multimodal retrieval evaluation. Note, we use the train-test split approach from (Shvetsova et al., 2022; Miech et al., 2019) for a fair comparison.

HowTo100M (Miech et al., 2019) serves as a pretraining dataset, encompassing instructional videos spanning 23,000 distinct activities and over 100 million samples, with a significant portion dedicated to cooking videos. Detailed annotations including textual descriptions and the inclusion of audio components further enhances its multimodal nature, paving the way for tasks like audio-visual retrieval.

YouCook2 (Zhou et al., 2018) aligns closely with the domain of the HowTo100M, primarily focusing on cooking related instructional content with 1 description per video. Although it serves as a benchmark for evaluating retrieval performance in comparison with works like (Shvetsova et al., 2022), it does not provide good insights on domain-agnostic capabilities of the model due to domain similarity with the pretraining dataset, HowTo100M.

MSR-VTT (Xu et al., 2016) is recognized as a benchmark for video retrieval and captioning, while it contrasts well with instructional datasets such as HowTo100M. The dataset encompasses 10,000 diverse video clips, from movie snippets to music and sports, each enriched with up to 20 human-generated descriptions. Different to the other two datasets, only approx 4% of MSR-VTT videos contain cooking and a total of 11% are instructional. Due to the diverse nature of its video contents, we consider the MSR-VTT dataset as a good representative for real-world retrieval scenarios. Following the earlier multimodal evaluations, we have used only those 968 videos that contain also audio.

4.2. Retrieval Setup and Metric

Consistent with previous studies (Rouditchenko et al., 2020; Nagrani et al., 2022; Shvetsova et al., 2022), it is assumed that there exists 1-to-1 correspondences between the query texts and the video files to be retrieved. This standard assumption means that in automatic performance evaluation only one video is considered as the correct retrieval result despite the fact that a human assessor could regard more than one video as correct. Correspondingly, the reported performance measures can be regarded as lower limits of the methods' human-observed performance.

The different ways of using trimodal data for TMR are defined and denoted as follows:

- $t \rightarrow v$: Videos are retrieved based on a text query. The representations of the text

and video modalities are both obtained from their respective unimodal experts. Audio is not used.

- $t \rightarrow v + a$: The text representation is obtained from the unimodal expert, whereas the video-audio fused feature is obtained as the element-wise sum of the respective unimodal experts.
- $t \rightarrow va$: Now, the fused feature is $f(\nu_i^{va}, \alpha_i^{va})$ and obtained via the modality pair va from the cross-attention block (see Eq. (3)).

Each mechanism focuses on visual-based retrieval, but the feature retrieved varies based on the type of interaction with audio.

Evaluation Metric. A commonly-used metric for video retrieval evaluations is the recall R@K, where K represents the number of top-ranked videos that are considered when measuring retrieval accuracy. The performance is reported for K in $\{1, 5, 10\}$ for gauging the model's recall across different retrieval depths. The higher the value of R@K, the more accurate the retrieval system is and the desired video is found within a smaller set of top retrieved videos.

4.3. Training Setting

We start our model training from the pre-trained weights of the Everything at Once (EAO) (Shvetsova et al., 2022) model. We keep the weights for layers which overlap with EAO and initialize the others from scratch. After this, we fine-tune the model with the task-specific MSR-VTT and YouCook2 data.

All the experiments are conducted based on the CLIP backbone (Radford et al., 2021) for the textual and visual representations. CLIP, pre-trained on the extensive Wikipedia-based image-text WiT dataset (Srinivasan et al., 2021), employs the ViT-B/32 model for its visual backbone and a BERT-like text encoder for its text backbone. This configuration extracts a 512-dimensional features for both video and text. The CLIP backbone is frozen with no updates while training the main architecture. For the audio, a trainable CNN is used as an audio backbone for a fair comparison to (Miech et al., 2019; Shvetsova et al., 2022). The audio backbone produces a 4096-dimensional feature per second.

In our architecture, the unimodal transformers and multimodal fusion transformer consist of only one block, also illustrated in Figure 2. The hidden size is set to 4096 with 64 attention heads. For projections of Eq. (3), the dimensions for the token and embedding space are 4096 and 6144, respectively (Shvetsova et al., 2022). For the loss function given in Eq. (5), the weights are set as $\lambda_{t,va} = \lambda_{t,tv} = \lambda_{t,ta} = 1$ with a temperature value of $\zeta = 0.05$. All the experiments are conducted for 25 epochs using the Adam optimizer (Kaiser et al., 2017) with a learning rate of $5 \cdot 10^{-5}$ and an exponential decay of 0.9. We have conducted our experiments using a computer featured by AMD Instinct GPUs and a node with eight parallel GPUs.

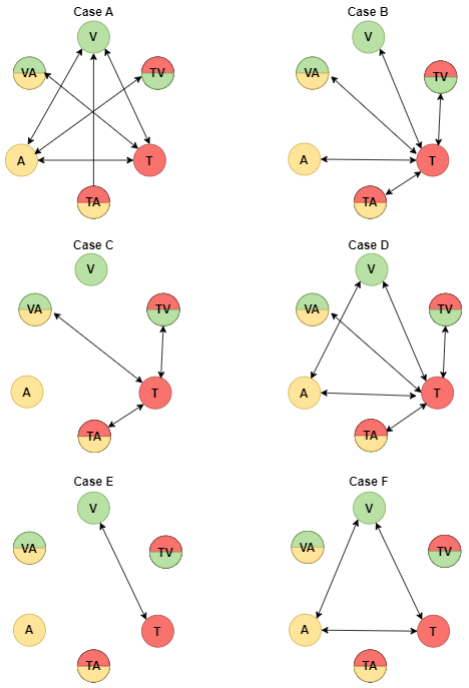

Figure 4: Loss variant Cases A to F. The color coding of the modalities follows that in Figure 2.

4.4. Loss Variants for Retrieval

The first aim of our study is to find the loss function that is optimal for performance in TMR tasks. For maintaining comparability with (Shvetsova et al., 2022), the experiments are conducted with pre-trained setting. We always employ bidirectional losses as shown in Eq. (4). However, the subtler challenge lies in identifying the bidirectional contrastive loss pairs that are the most effective for TMR. Figure 4 showcases six alternative combinations, labeled A to F, of loss functions for Eq. (5) that we studied.

Table 1 reports the retrieval results in the fine-tuning setting with the MSR-VTT data on the proposed and the original EAO model, which introduced only the loss Case A. For the $t \rightarrow v$ retrieval mechanism, Case E loss results in the highest performance among all cases. However, for the $t \rightarrow v + a$ mechanism, Case C and E losses compete, and for the $t \rightarrow va$ mechanism, Case C loss outperforms others. Overall, we achieve the best retrieval performance using the combination of $t \rightarrow va$ and Case C loss, both for the proposed and the original EAO model.

Retrieval	Case	EAO Arch.			Ours		
		R@1	R@5	R@10	R@1	R@5	R@10
$t \rightarrow v$	A	28.2	58.7	68.4	27.5	59.8	69.7
	B	33.4	63.2	74.2	31.2	63.9	**74.3**
	C	27.1	57.1	69.7	28.4	57.7	71.1
	D	29.6	60.5	72.3	24.3	54.2	66.5
	E	**35.5**	63.9	74.2	34.8	**64.3**	73.6
	F	32.9	63.4	73.7	30.5	60.5	72.0
$t \rightarrow v + a$	A	26.2	56.8	67.4	25.4	57.4	65.4
	B	31.1	60.6	72.6	26.4	56.4	70.8
	C	28.0	58.7	70.7	34.9	**66.0**	**76.1**
	D	30.1	59.6	71.0	22.0	51.2	64.6
	E	**35.3**	64.4	74.8	35.1	63.8	74.3
	F	28.5	58.4	71.4	26.2	56.4	70.1
$t \rightarrow va$	A	31.8	62.4	74.3	32.4	60.4	69.8
	B	37.3	66.5	**77.9**	33.1	64.1	77.1
	C	**38.3**	**67.3**	77.6	36.8	**67.3**	77.6
	D	36.9	65.7	76.9	27.0	57.2	70.4
	E	34.6	63.5	73.8	33.4	62.5	72.4
	F	22.8	52.3	66.0	15.8	38.7	52.3

Table 1: Results on MSR-VTT using the original EAO and our architectures with various loss variants (shown in Figure 4) in different TMR retrieval mechanisms (i.e., $t \rightarrow v$, $t \rightarrow v + a$, $t \rightarrow va$). The grayed results are the only ones reported in the original EAO work. Note, our architecture uses partial weights, while EAO uses all weights pretrained on HowTo100M.

Retrieval	Architecture	R@1	R@5	R@10
$t \rightarrow v$	3×Umt	**31.6**	**61.4**	**73.8**
	1×Caft	29.1	60.5	71.5
	3×Umt+3×Caft	22.7	40.9	53.7
	EAO	27.1	57.1	69.7
	1×Umt+1×Caft	28.8	57.4	69.0
	3×Umt+1×Caft (Ours)	28.6	57.4	70.5
$t \rightarrow v + a$	3×Umt	33.5	62.4	74.9
	1×Caft	31.5	61.4	72.6
	3×Umt+3×Caft	29.5	61.9	73.3
	EAO	28.0	58.0	70.7
	1×Umt+1×Caft	33.6	63.6	74.6
	3×Umt+1×Caft (Ours)	**34.9**	**66.0**	**76.1**
$t \rightarrow va$	1×Caft	34.7	63.8	74.5
	3×Umt+3×Caft	30.3	61.4	74.8
	EAO	**38.3**	**67.3**	**77.6**
	1×Umt+1×Caft	35.2	66.8	76.7
	3×Umt+1×Caft (Ours)	36.8	**67.3**	**77.6**

Table 2: Comparison between architectural variants using loss Case C on the MSR-VTT dataset. *Umt* stands for unimodal transformers and *Caft* stands for cross-attention fusion transformer.

We can see that, by strategically leveraging contrastive loss, Case C loss effectively increases mutual information between text and other modalities, correlating them more cohesively in the TMR task. This is especially clear when compared with the EAO model's original Case A loss.

4.5. Architectural Variants for Retrieval

In addition to our main architecture depicted in Figure 2, we also explored a number of other architectures for TMR tasks. The variants rely on various transformer-block combinations, including unimodal transformers, i.e., *Umt*, and cross-attention fusion transformers, i.e., *Caft*, to identify the most effective architecture. Table 2 summarizes the performance of each architecture trained with Case C loss on the MSR-VTT dataset.

Notably, our model consistently outperforms other architectural variants across all retrieval scenarios. Our model outperforms the EAO architecture in Case C loss, excelling in R@1, R@5, and R@10 for the $\rightarrow v + a$ mechanism. Meanwhile, it maintains comparable performance in R@5 and R@10 for $t \rightarrow va$. It is worth noting that our model uses partial weights from EAO pretrained on the HowTo100M and a fully pretrained model is expected to achieve better performance. Moreover, unlike the fusion transformer we designed, the fusion transformer designed by the EAO grows exponentially as the number of modalities increases.

Additionally, the architectural variant with a single *Umt* and *Caft*, i.e., 1×*Umt*+1×*Caft*, is the closest competitor to our model. While our model consistently outperforms this variant across all retrieval tasks, the difference in performance is relatively marginal. This observation suggests that the 1×*Umt*+1×*Caft* combination could be a strong contender, especially in scenarios where there is a need to reduce the number of model parameters.

4.6. Comparison with State-of-the-Art

Table 3 shows a comparison of our proposed model with a number of other TMR models pretrained on the HowTo100M data and tested on the MSR-VTT and YouCook2 datasets. It can be seen that our model consistently outperforms existing methods, emphasizing its prowess in multimodal retrieval tasks. Specifically, when juxtaposed with the EAO* (replicated EAO) results, our approach demonstrates significant advancements in both the $t \rightarrow v + a$ and $t \rightarrow va$ retrieval tasks. For the $t \rightarrow v + a$ task, our model on MSR-VTT demonstrated a remarkable 33.2% relative improvement in R@1, a 16.2% rise in R@5, and a 12.9% boost in R@10. Similarly, for the $t \rightarrow va$ task, we noticed a 15.7% increase in R@1, a 7.9% enhancement in R@5, and a 4.4% growth in R@10. These results not only underscore the effectiveness of our fusion style, architecture, and loss selection, but also solidify our model's superiority in the TMR task.

Table 3 also shows the impact of the pretraining dataset on the performance. HowTo100M focuses on human speech and has less diverse audio, with textual descriptions from ASR and a

Method	Pretrain. Data	Retrieval	YouCook2			MSR-VTT		
			R@1	R@5	R@10	R@1	R@5	R@10
VAST (Chen et al., 2023b)	VAST27M	$t \rightarrow v + a$	50.4	74.3	80.8	63.9	84.3	89.6
VALOR (Chen et al., 2023a)	VALOR1M	$t \rightarrow v + a$	–	–	–	54.4	79.8	87.6
LAV (Nagrani et al., 2022)	VideoCC3M	$t \rightarrow v + a$	–	–	–	35.8	65.1	76.9
AVL (Rouditchenko et al., 2020)	HowTo100M	$t \rightarrow v + a$	30.2	55.5	66.5	22.5	50.5	64.1
EAO*	HowTo100M	$t \rightarrow v + a$	29.7	58.6	69.4	26.2	56.8	67.4
LAV (Nagrani et al., 2022)	HowTo100M	$t \rightarrow v + a$	–	–	–	33.1	62.3	72.3
Ours	HowTo100M	$t \rightarrow v + a$	**32.7**	**63.7**	**74.3**	**34.9**	**66.0**	**76.1**
EAO (Shvetsova et al., 2022)	HowTo100M	$t \rightarrow va$	–	62.7	75.0	–	62.1	72.9
EAO*	HowTo100M	$t \rightarrow va$	32.3	62.1	72.9	31.8	62.4	74.3
Ours	HowTo100M	$t \rightarrow va$	**34.8**	**64.2**	**75.6**	**36.8**	**67.3**	**77.6**

Table 3: Comparison with state-of-the-art works which used audio with video for the text-to-multimodal retrieval. EAO shows the results reported in (Shvetsova et al., 2022) whereas EAO* shows our replicated results. Note, VAST, VALOR and LAV results are shown for the completeness of the study as the differences in the pretraining prevent direct comparisons of the results.

domain-specific emphasis on cooking videos. In contrast, smaller datasets like VAST27M (Chen et al., 2023b), VALOR1M (Chen et al., 2023a), and VideoCC3M (Nagrani et al., 2022) perform very well, most probably due to their better quality and diversity.

4.7. Analysis on Query Length

In this section, we assess the impact of text query length on the efficacy of retrieval systems. In addition to the default test query set - MSR-VTT 1k used by (Shvetsova et al., 2022; Miech et al., 2019), we build two additional query sets, namely "short" and "long", that vary in terms of the query text length. The distributions of query lengths are depicted in Figure 5. A notable discrepancy exists between the recommended evaluation query set and the range of short to long queries present in the dataset.

The results in Table 4 reveal a compelling trend: as the query becomes lengthier and more detailed, the system's proficiency in pinpointing the correct video improves, evidenced by a significant 16% enhancement in the R@1 metric, when compared to the MSR-VTT 1k results. This underscores the pivotal role of text query length in influencing performance outcomes. Consequently, it raises a pertinent question regarding the reliability of MSR-VTT as a benchmark for gauging retrieval system performance. The experiment reinforces the hypothesis presented in (Rodriguez et al., 2022). Their approaches provide a promising direction for future work, potentially leading to the design of improved TMR benchmarks. This could be of significant importance in the field.

4.8. Qualitative Results with Audio

To show the effectiveness of including the audio modality, we compare retrieval results between the $t \rightarrow v$ and $t \rightarrow va$ mechanisms. Notably, from 968 videos (the number of videos with audio in MSR-VTT test set), the $t \rightarrow va$ mechanism retrieved 44 more R@1-correct videos than $t \rightarrow v$ in the test set. Figure 6 provides examples that appeared only in the $t \rightarrow va$ retrieval results. On inspecting these videos, the benefits of audio integration are in some cases evident whereas some other cases appear more coincidental. In particular, even though the speech of the man in the first clip is not explicitly recognized, the overall multimodal context is hinting towards invitation. Also, in the cases of the video game and dancing videos, the audible modality is indeed strongly supporting the retrieval. On the other hand, both the chair and duck examples are correct retrievals, but the audible contents cannot be credited for that.

Retrieval	Query Length	R@1	R@5	R@10
$t \rightarrow v$	Short	17.3	41.9	54.9
	MSR-VTT 1k	28.6	57.4	70.5
	Long	**37.5**	**69.1**	**80.6**
$t \rightarrow v + a$	Short	21.0	46.9	58.5
	MSR-VTT 1k	34.9	66.0	76.1
	Long	**42.3**	**73.6**	**83.6**
$t \rightarrow va$	Short	21.3	47.3	58.9
	MSR-VTT 1k	36.8	67.3	77.6
	Long	**42.7**	**73.4**	**83.7**

Table 4: Query length vs. TMR performances on the MSR-VTT dataset with our proposed architecture and loss Case C.

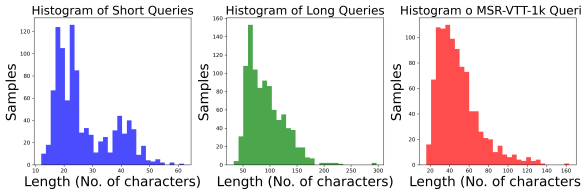

Figure 5: Histograms of query lengths of the three test query sets from MSR-VTT text descriptions.

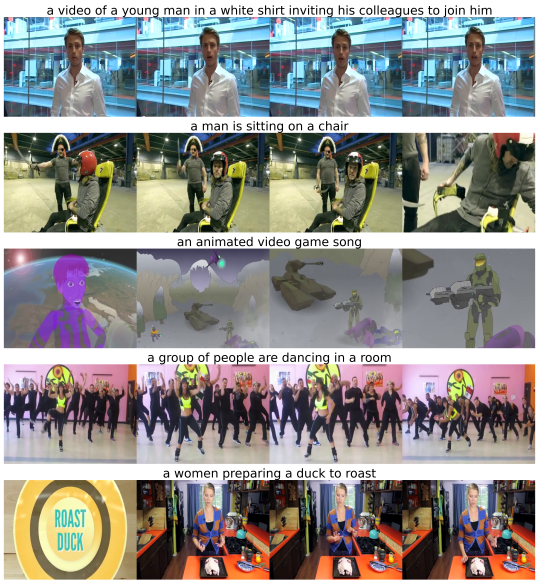

Figure 6: Examples retrieved with $t \rightarrow va$ but not with $t \rightarrow v$ on the MSR-VTT-1k test set.

5. Discussion

Many recent studies have utilized vision-language models based on self- or cross-attention mechanisms to enhance downstream tasks, including text-to-video retrieval. However, video data contains not only visual but also audio information. Moreover, the integration of even more modalities (e.g., haptics, 3D) into transformer architectures for retrieval tasks still remains as a very seldom addressed research question.

Specifically, there are a few recent works that incorporate the audio modality into multimodal transformer-based architectures (see Table 3). While these models (e.g., the EAO model) already face increased data complexity due to the additional audio modality, the models require further design considerations in terms of multimodal fusion with transformers. Overall, this introduces new research questions and applications, where multiple modalities can be efficiently utilized and fused to enrich representations in shared multimodal embedding spaces.

Our primary motivation has been to introduce a scalable and flexible text-to-multimodal retrieval (TMR) architecture, open for future expansion with other modalities. The cross-attention mechanism used in our model is the same for every modality. One can further extend our model by incorporating state-of-the-art architectures in the backbones and adding processing steps with minimal changes. Covering a new modality requires adding an unimodal transformer with an effective loss combination to the existing objective function, without training the model from scratch. Each modality's unique pipeline allows us to tailor the obtained shared embedding space for interaction of modalities in the multimodal transformer. The results of our experiments have shown that, in addition to architectural advantages, our architecture is on par with the latest technology studies.

One can finally observe from the results in Table 3 that our model performed even better on the MSR-VTT dataset than on YouCook2 despite the larger domain gap from the HowTo100M pretraining dataset. We can thus conclude that the discrepancy between the different domains of the pretraining and fine-tuning datasets does not seem to negatively affect the effectiveness of our model.

The information embedded within the textual contents plays a pivotal role in multimodal retrieval tasks with text queries. Particularly, the quantity and quality of these queries significantly influence the performance of applications. We have analyzed the impact of query length, showing that longer descriptions benefit the retrieval task (see Table 4). This highlights the importance of future work, such as the integration with large language models (LLMs), to enrich text queries.

Multimodal retrieval assessments often prioritize the text-to-video retrieval task, which is crucial for efficient web and personal media searches. At the same time, the challenge of integrating text with other modalities through cross-modal representations is central to the design of many multimodal models. Therefore, our study has aimed to improve text-to-multimodal retrieval of videos by leveraging all available data modalities, i.e., text, audio and visual, while maintaining their integrity. Our primary focus has been on retrieval, but our architecture is versatile and could be applied to other text-related tasks involving multiple modalities, such as captioning and visual question answering.

6. Conclusions and Future Work

In this work, we proposed a novel hierarchical approach for the text-to-multimodal retrieval task. Our model architecture is based on the mid-fusion strategy and is easily applicable and scalable for other tasks that require multimodal interaction. Our work also bolsters the concept of task-specific fine-tuning to tailor the generic multimodal representation space for text-to-multimodal retrieval. Targeted combination of loss functions significantly enhanced task performance compared to using a non-targeted combination. One persistent limitation is the impact of pretraining dataset selection, which can severely affect the performance in all multimodal tasks. Exploiting large multimodal domain-agnostic pretraining datasets holds immense potential for the improvement and further applicability of our model in real-world scenarios.

7. Acknowledgements

This research has been funded by the Research Council of Finland in project #345791 *Understanding speech and scene with ears and eyes (USSEE)*. We acknowledge CSC – IT Center for Science, Finland for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through its Extreme Scale Access program.

8. Bibliographical References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6836–6846.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems*, 29.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. 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 1728–1738.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. 2021. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8012–8021.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023a. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023b. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *arXiv preprint arXiv:2305.18500*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. 2019. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3965–3969. IEEE.

Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. 2011. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):797–819.

Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mohamed Omar. 2023. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12054–12064.

Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):1–19.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

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

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. 2022. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Guoping Qiu. 2022. Challenges and opportunities of image and video retrieval. *Frontiers in Imaging*, 1.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Pedro Rodriguez, Mahmoud Azab, Becka Silvert, Renato Sanchez, Linzy Labson, Hardik Shah, and Seungwhan Moon. 2022. Fighting fire with fire: Assessing the validity of text-to-video retrieval benchmarks. *arXiv preprint arXiv:2210.05038*.

Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. 2020. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. 2022. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20020–20029.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

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

Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 970–981.

Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu. 2023. Deep learning for video-text retrieval: a review. *arXiv preprint arXiv:2302.12552*.

9. Language Resource References

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2630–2640.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. 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.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.