Rethinking the Multimodal Correlation of Multimodal Sequential Learning via Generalizable Attentional Results Alignment

Tao Jin[1,2] Wang Lin[1] Ye Wang[1] Linjun Li[1] Xize Cheng[1] Zhou Zhao[1,2†]
[1]Zhejiang University
[2]Shanghai AI Laboratory
*

Abstract

Transformer-based methods have gone mainstream in multimodal sequential learning. The intra and inter modality interactions are captured by the query-key associations of multi-head attention. In this way, the calculated multimodal contexts (attentional results) are expected to be relevant to the query modality. However, in existing literature, the alignment degree between different calculated attentional results of the same query are under-explored. Based on this concern, we propose a new constrained scheme called Multimodal Contextual Contrast (MCC), which could align the multiple attentional results from both local and global perspectives, making the information capture more efficient. Concretely, the calculated attentional results of different modalities are mapped into a common feature space, those attentional vectors with the same query are considered as a positive group and the remaining sets are negative. From local perspective, we sample the negative groups for a positive group by randomly changing the sequential step of one specific context and keeping the other stay the same. From coarse global perspective, we divide all the contextual groups into two sets (i.e., aligned and unaligned), making the total score of aligned group relatively large. We extend the vectorial inner product operation for more input and calculate the aligned score for each multimodal group. Considering that the computational complexity scales exponentially to the number of modalities, we adopt stochastic expectation approximation (SEA) for the real process. The extensive experimental results on several tasks reveal the effectiveness of our contributions.

1 Introduction

Multimodal sequential learning, which aims to process and understand the semantic information from multiple modalities (e.g., vision, language, audio) with machine learning skills, has drawn increasing attention recently. Many endeavors (Gabeur et al., 2020; Pham et al., 2019; Zadeh et al., 2018a,b) are devoted to the design of multimodal interactive mode and effective individual representation learning. Transformer-based multimodal interaction methods (Gabeur et al., 2020; Tsai et al., 2019) occupy the mainstream position in multimodal interaction area. Compared with vanilla methods, Transformer-based methods could achieve relatively superior performances with deep stacked attention blocks (Vaswani et al., 2017) and a suitable number of training samples. Concretely, by treating one modality as query (e.g., text) and the other modalities as keys (e.g., visual, audio), the multimodal contextual sequences (attentional results) can be obtained by query-key associated mechanism. In this way, the calculated attentional results are expected to be relevant to the query modality. However, the alignment degree between different calculated attentional results of the same query are under-explored. Specifically, if we have three modalities (i.e., text, visual, audio), we employ Transformer to correlate text-visual and text-audio interaction respectively. As shown in Fig. 1, when the text query is "rather", the attentional

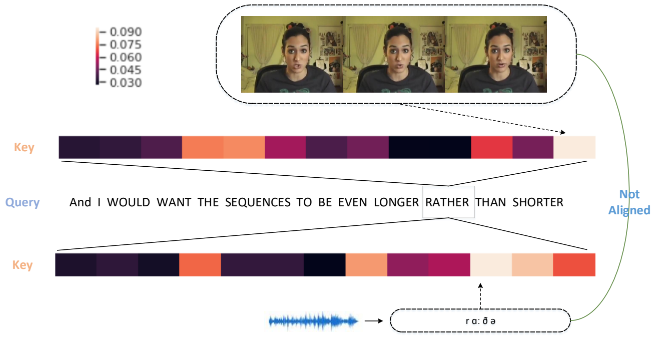

0.090
0.075
0.060
0.045
0.030
Key
Query
And I WOULD WANT THE SEQUENCES TO BE EVEN LONGER RATHER THAN SHORTER
Key
Not Aligned
rɑ:ðə

Figure 1: An example of calculated attentional results, where the colours denote attention weights and we present the text, visual, audio content of a speaker. The text query is "rather", the visual result ("shorter") is not aligned with audio result ("rather").

† denotes corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5247–5265
August 11-16, 2024 ©2024 Association for Computational Linguistics

results (“shorter”) of text-visual pair has obvious problems, while the audio localization (“rather”) is correct. During training, the mistakes of query-key association are inevitable. At this time, constraining the alignment of audio and visual attentional results can polish the capture mistakes indirectly.

Motivated by the observations, in this paper, we propose a new constrained strategy called Multimodal Contextual Contrast (MCC), which could align multimodal attentional results from both local and global perspectives, making the attentional capture more efficient. Specifically, the multimodal attentional results of different modalities are calculated with multi-head attention first and then mapped into a common feature space. The sequence lengths of the multimodal attentional results are same as that of query modality, we denote the multiple attentional sequences as $c_i \in \mathbb{R}^{t_a \times d} (i \in [n])$, where d is the feature dimension and n is the number of modalities. Those attentional vectors with the same query (at the same sequential step $t \in [t_a]$) are considered as a positive group and the remaining groups (at least one of the attentional vectors is with different query) are negative. From the local perspective, we sample the negative groups for a positive group by randomly changing the sequential step of one specific context vector and keeping the other stay the same. Totally, the number of negative groups is $n(t_a - 1)$ for a positive group. From the global perspective, we divide all the groups into two sets (i.e., aligned and unaligned), making the total score of aligned groups relatively large. For the implementation of contrastive constraints, we extend the vectorial inner product operation for more input and compute the aligned score for each multimodal group. Considering that the computational complexity of relevance scores scales exponentially to the number of modalities, we adopt stochastic expectation approximation for the real process. We conduct extensive experiments on three tasks, the experimental results show that MCC could achieve competitive results compared with the state-of-the-art methods. To sum up, the contributions of our work are four-folded:

- We propose a constrained strategy called Multimodal Contextual Contrast (MCC) for multimodal sequential learning, which conducts contrastive constraints for the multiple calculated attentional results. To the best of our knowledge, it is the first time to conduct contrastive scheme for the calculated attentional results.
- We develop the contrastive mechanism from both fine-grained local and coarse-grained global perspectives, making the attentional capture more accurate indirectly.
- Considering that the computational complexity of relevance scores scales exponentially to the number of modalities, we adopt stochastic expectation approximation (SEA) for the real process.
- We conduct extensive experiments on three tasks, multimodal sentiment analysis (with video, audio, text modalities), speaker traits recognition (with video, audio, text modalities), and video retrieval (with motion, scene, OCR, audio, speech, face, appearance modalities). The experimental results show that MCC could achieve competitive results compared with state-of-the-art methods.

2 Related Work

Multimodal Interaction. Existing multimodal interaction methods could be categorized into Transformer-based and Non-Transformer-based methods. As for the former, Zadeh et al. (2016b) proposes to train the model on simply concatenated multimodal features for prediction. Poria et al. (2017) correlates multiple modalities with a context-dependent fusion method. Zadeh et al. (2018b) explicitly accounts for both interactions in a neural architecture and continuously models them through time. As for the latter, Tsai et al. (2019) proposes multimodal transformer to boost the interactions between multiple modalities. LMF-MulT (Sahay et al., 2020) combines the LMF and MulT to process multimodal sequential information, First LMF-MulT aligns different modalities at sequential-step level, fusing different modalities. The results are input to the Transformer module. Li et al. (2022) proposes the modal-order-aware network to integrate the three modalities in a certain order to distinguish the importance of different modalities. Yang et al. (2023) conducts contrastive feature decomposition from the sample level.

Contrastive Learning. According to the modality of data, existing methods can be divided into two categories, i.e., single-modality based and multi-modality based contrastive learning.

Considering the single-modality based methods, Wu et al. (2018) uses a memory bank which stores previously-computed representations and noise-contrastive estimation (NCE) (Gutmann and Hyvärinen, 2010) to tackle the computational challenges imposed by the large number of instance classes. MoCo (He et al., 2020) further improves such a scheme by storing representations from a momentum encoder in dynamic dictionary with a queue. As for the multi-modality based methods (Hager et al., 2023; Mustafa et al., 2022; Yang et al., 2023), the common strategy is to explore the natural correspondences among different modalities and use contrastive learning to learn representations by pushing modalities describing the same scene closer, while pushing modalities of different scenes apart. However, different from the existing multimodal contrastive strategies, we consider the problem from the view of attentional results and their time steps, instead of input features.

3 Contrastive Contextual Alignment

In this section, we would introduce MCC with the following scheme. First, we give simple illustration of the definition and background of some symbols. Second, we would introduce the local contrastive alignments and global contrastive alignments in detail (as shown in Fig. 2). Third, considering the complexity, we adopt the theoretical approximation for computation reduction. To be more intuitive, we further provide the complexity analysis of vanilla MCC and Improved MCC.

3.1 Multimodal Contextual Sequences

Suppose that there exists n modalities with sequential representation $v_i \in \mathbb{R}^{t_i \times d}, i \in [n]$, where t_i denotes the sequence length of i-th modality, d denotes the feature dimension of all the modalities. For convenience, we choose one modality $v_a \in \mathbb{R}^{t_a \times d}$ as "anchor (query)". As we know, there are two forms of interactions among different modalities: modality-specific interactions and cross-modal interactions. Considering the former interactions, we treat the v_a as "key (value)". Considering the latter interactions, we treat $v_i (i \neq a)$ as "keys (values)". In this paper, we employ the mainstream Transformer structure for multimodal sequential learning. Specifically, the modality-specific and cross-modal interactions of v_a can be expresses as follows:

$$\begin{aligned} c_a &= \text{Self\_ATT}(v_a) \\ c_i &= \text{Cross\_ATT}(v_a, v_i) \end{aligned} \tag{1}$$

where $c_a \in \mathbb{R}^{t_a \times d}$ denotes the modality-specific attentional results and $c_i \in \mathbb{R}^{t_a \times d}$ denotes the cross-modal attentional results. The only difference between functions "Self_ATT" and "Cross_ATT" is the key (value). To be more intuitive, we can further rewrite these two functions as follows:

$$\begin{aligned} \text{Self\_ATT}(v_a) =& \text{ATT}(v_a, v_a, v_a) \\ \text{Cross\_ATT}(v_a, v_i) =& \text{ATT}(v_a, v_i, v_i) \end{aligned} \tag{2}$$

where "ATT" denotes the multi-head attention mechanism (Vaswani et al., 2017) which is widely used in computer vision/natural language processing/multimodal analysis community. In the following sections, we merge the contexts c_a and $c_i (i \neq a)$ and employ $c_i \in \mathbb{R}^{t_a \times d} (i \in [n])$ for illustration. In practice, we implement multiple projection layers following the attentional sequences. For convenience, we still utilize d to denote the common feature dimension.

3.2 Local Contrastive Alignments

The basic of multi-head attention mechanism is the inner-product operation for query-key similarity, while most of the existing methods do not consider the relevance of different contextual sequences of the same query. For example, the context (attentional) vectors corresponding to the t-th time step ($v_a^t \in \mathbb{R}^d$) of query modality are $c_i^t \in \mathbb{R}^d$ where $i \in [n]$. According to the calculation rules of inner-product operation, each c_i^t would be related to v_a^t. However, the alignments between the context vectors ($c_i^t, i \in [n]$) are not strictly evaluated, which may influence the attentional capture. Inspired by the widely-studied contrastive learning techniques, we divide the groups of context vectors into positive sets and negative sets. Note each context group contains n vectors correspond to n random sequential steps of n modalities, thus, the number of total groups is $(t_a)^n$. We facilitate the contrasts from two perspectives: local alignments and global alignments, which are complementary to each other. In this section, we introduce the local contrastive alignments in detail. Following the illustration above, the context vectors of the same query (at the same step) are relevant and treated as positive (totally t_a groups). The main challenge

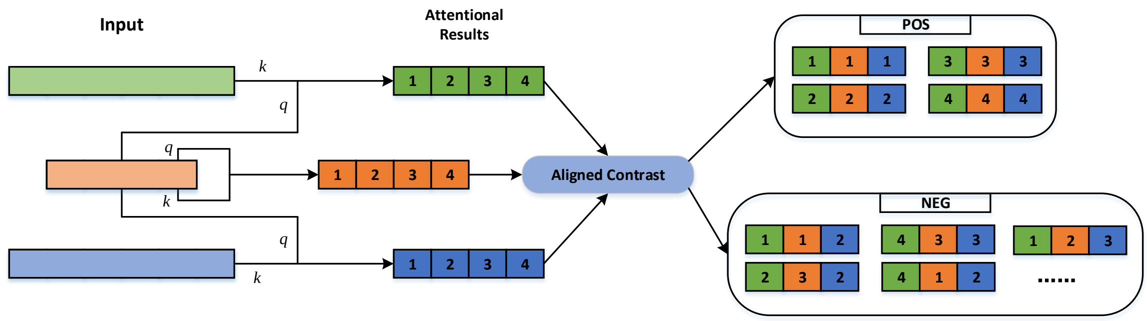


Figure 2: The overall framework of MCC, where different colors denote different modalities, k and q denote key and query, the specific numbers denote the attentional results of the specific query vectors at corresponding sequential steps. We divide the contextual groups into positive (top) and negative (bottom) sets for contrastive learning. Further details are shown in Fig. 3, we show the difference of local and global contrastive constraints.

is to sample some negative groups. We start by introducing the alignments for two contextual sequences and then employ the conclusion for the generalization of more contextual sequences.

3.2.1 The Alignments for Two Contextual Sequences

Suppose that there are two contextual sequences c_1 and $c_2 \in \mathbb{R}^{t_a \times d}$. The primary objective of the training is to maximize the alignment degree between the positive pairs (i.e., c_1^t and c_2^t). Thus, we first define the alignment function by using the normalized inner product as:

$$\mathrm{A}(c_1^t, c_2^t) = \frac{\langle c_1^t, c_2^t \rangle}{||c_1^t||||c_2^t||} \tag{3}$$

where $\langle , \rangle$ denotes the inner product operation. The range of function A() is $[-1, 1]$. We hope that the scores of positive pairs are close to 1. However, merely optimizing the alignment of positive pairs ignores the important positive-negative relation knowledge (Mikolov et al., 2013b). To make the training process more informative, we reform the overall objective in the contrastive learning manner (Arora et al., 2019; Van den Oord et al., 2018) with Noise Contrastive Estimation (NCE) loss (Mnih and Teh, 2012; Mikolov et al., 2013b). Specifically, we consider the fact that one context vector is more related to the context vector with the same query (at the same sequential step among all the steps). Then, we can formulate the overall NCE objective as follows:

$$\begin{aligned} \mathcal{L}_l &= -\sum_{t=1}^{t_a} \log \left[\frac{\exp\left(\mathrm{A}(c_1^t, c_2^t)\right)}{\exp\left(\mathrm{A}(c_1^t, c_2^t)\right) + \mu}\right] \\ \mu &= \sum_{t'=1, t' \neq t}^{t_a} \left(\exp\left(\mathrm{A}(c_1^t, c_2^{t'})\right) + \exp\left(\mathrm{A}(c_1^{t'}, c_2^t)\right)\right) \end{aligned} \tag{4}$$

where $\mathrm{A}(c_1^t, c_2^t)$ denotes the positive pair of the query at t-th sequential step, $\mathrm{A}(c_1^t, c_2^{t'})$ and $\mathrm{A}(c_1^{t'}, c_2^t)$ denote the negative pairs according to the natural facts stated above. Such objective in Eq. 4 explicitly encourages the alignment of positive pair while separates the negative pairs.

3.2.2 Generalization for More Contextual Sequences

With the conclusion for two contextual sequences, we discuss the condition of more contextual sequences. One simple idea is to treat the contextual sequences as multiple two-sequence pairs and utilize the existing conclusion of Eq. 4. However, this way neglects the correlation among all the modalities. Thus, we consider the contrastive constraints from a more general perspective by jointly processing all the contextual sequences. Based on one specific positive group $\{c_i^t \in \mathbb{R}^d | i \in [n]\}$ of the same query at t-th sequential step, we try to change some sequential steps of the attentional vectors and analyze the relative correlation of all the records. After the multiple replacements, we obtain two conclusions.

We first define the relevance scores for the n-vector sets based on those of two-vector pairs. The detailed process is shown as follows:

$$\mathrm{A}(\{c_i^t | i \in [n]\}) = \sum_{i=1}^{n} \sum_{j=1, j<i}^{n} \mathrm{A}(c_i^t, c_j^t) \tag{5}$$

Proposition 1: Suppose that we have the group of contexts for the t-th sequential step $c_i^t \in \mathbb{R}^d$ where $i \in [n]$. When we randomly change the sequential step of one specific context, the obtained new group has less relevance than the original positive group. Further, if we treat the score of positive

n-vector group consisting of the scores of multiple ($\frac{n(n-1)}{2}$) positive two-vector pairs like Eq. 5, the scores of negative n-vector groups (only changing one specific context) also contain the scores of all the negative two-vector pairs. **The detailed analysis is in the appendix (Sec. B).**

Proposition 2: Only changing the sequential step for the query vector of one modality can be considered as hard negative mining. Following the nature of contrastive learning (Robinson et al., 2020), to enhance the alignment of the positive set, we can improve the difficulty of the negative set. Based on a specific positive set, if we change the step of one modality, the number of irrelevant two-vector pairs increases by $O(n)$. When we change the steps of s modalities, the complexity is $O(sn)$.

Based on the above observations, we create the negative groups for the specific positive groups. Concretely, we sample the negative groups with only one different sequential step. Totally, the number of the negative groups is $n(t_a-1)$ for a specific postive group. The local contrastive constraints can be expressed as follows:

$$\mathcal{L}_l = -\sum_{t=1}^{t_a} \log\left[\frac{\exp\left(\mathrm{A}(\{c_i^t|i\in[n]\})\right)}{\exp\left(\mathrm{A}(\{c_i^t|i\in[n]\})\right)+\mu}\right] \qquad (6)$$
$$\mu = \sum_{*} \exp\left(\mathrm{A}(\{c_i^{t'_i}|i\in[n]\})\right)$$

where $*$ denotes the sampling condition of negative groups (i.e., only one of $\{t'_i|i\in[n]\}$ is not equal to t, the others are equal to t). Although the local contrastive constraints can align the multi-modal contexts at a fine granularity, most of the negative sets are discarded. Relying on this loss function alone, we may obtain a locally optimal solution. To solve this concern, we propose a complementary contrastive constraint, which align the context vectors from a coarse global perspective.

3.3 Global Contrastive Alignments

To make full use of the negative context (attentional results) sets, we propose a complementary global contrastive strategy. Specifically, we divide all the context groups into two sets, one includes all the positive groups and the other includes all the negative groups. From a coarse-grained perspective, the scores of positive groups should be close to $\frac{n(n-1)}{2}$ and the scores of negative sets are less than $\frac{n(n-1)}{2}$. The optimization goal is to make the more relevant set dominate, which fits our intuition. The formula expression is shown as follows:

$$\mathcal{L}_g = -\log\left[\frac{\sum_{t=1}^{t_a}\exp\left(\mathrm{A}(\{c_i^t|i\in[n]\})\right)}{\sum_{t=1}^{t_a}\exp\left(\mathrm{A}(\{c_i^t|i\in[n]\})\right)+\mu}\right] \qquad (7)$$
$$\mu = \sum_{*} \exp\left(\mathrm{A}(\{c_i^{t'_i}|i\in[n]\})\right)$$

where $*$ denotes the sampling condition of negative groups (i.e., not all of $\{t'_i|i\in[n]\}$ are equal), $\mathcal{L}_g$ denotes the global contrastive loss, which can control the relative distributions of positive relevance and negative relevance, to some extent.

3.4 Stochastic Expectation Approximation

When the number of modalities increases, the computational complexity would become unexpectedly large. Partial complexity comes from combinational summation operation for relevance score. Inspired by the kernel approximation skills (Kar and Karnick, 2012), we develop an efficient method called Stochastic Expectation Approximation (SEA) to calculate the relevance score. Following the assumptions above, the exp-based relevance score of a context group is expressed as:

$$\exp\left(\sum_{i=1}^{n}\sum_{j=1,j<i}^{n}\mathrm{A}(c_i^t,c_j^t)\right)$$
$$=\exp\left(\sum_{i=1}^{n}\sum_{j=1,j<i}^{n}\langle\frac{c_i^t}{||c_i^t||},\frac{c_j^t}{||c_j^t||}\rangle\right) \qquad (8)$$

The main challenges of the approximation are two-folded: First, we should reduce the square-level complexity. Second, we should consider the reconstruction of non-linear function $\exp()$. The Stochastic Expectation Approximation is the extension of binary kernel reconstruction, which can be expressed as follows:

$$\exp\left(\sum_{i=1}^{n}\sum_{j=1,j<i}^{n}\langle v_i,v_j\rangle\right) =$$
$$\mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\prod_{k=1}^{n}\exp\left(w^\top v_k - \frac{\|v_k\|^2}{2}\right)\right] \qquad (9)$$

where we utilize v_i to denote $\frac{c_i^t}{||c_i^t||}$ for convenience. $\mathbb{E}$ and $\mathcal{N}(0,\mathbf{I}_d)$ denote expectation operation of different random features w and the sampling distribution of w, respectively. **The detailed derivations can be found in the appendix (Sec. C).** Note that we also provide an analysis of the trade-off between the sampling number (computational complexity) and performance in the appendix Eq. 18 and Sec. D.

3.5 Complexity Analysis

In this section, we detailedly analyze the alignment complexity before and after the approximation. We divide the analysis into two parts, global contrastive alignments and local contrastive alignments. The total complexity of the global contrastive alignments is exponential $O(n^2(t_a)^n)$, the square complexity $O(n^2)$ arises from the combinational addition operations and the exponential complexity $O((t_a)^n)$ arises from a large number of context sets. We argue that the SEA approximation can reduce the square complexity to a linear level $O(n)$. Specifically, the relevance score of a set can be calculated with continuous multiplication operation according to the Eq. 9. As for the exponential part, we calculate the sum of all the relevance scores like $\sum_i \sum_j \sum_k a_i b_j c_k = (\sum_i a_i)(\sum_j b_j)(\sum_k c_k)$ (from exponential $O((t_a)^n)$ to linear $O(nt_a)$). **The detailed analysis is shown in the appendix (Sec. E).** Therefore, the complexity with approximation changes from $O(n^2(t_a)^n)$ to $O(n^2 t_a)$.

The total complexity of the local contrastive alignments is $O(n^3(t_a)^2)$, where the square term $O(n^2)$ also arises from the combinational addition operation, the term $O(n(t_a)^2)$ arises from the number of negative sets in the local contrastive constraints. This term can be easily reduced to $O(nt_a)$ with vanilla summation for multiple sequential steps. Thus, the complexity is also $O(n^2 t_a)$. With the SEA approximation, the total computational complexity changes from $O(n^3(t_a)^2)$ to $O(n^2 t_a)$.

3.6 Training

We evaluate MCC on three tasks, including multimodal sentiment analysis, speaker traits recognition, and video retrieval. MCC is treated as an auxiliary constraint for these tasks. Suppose that $\mathcal{L}_t$ denotes the loss of the original task, the final optimization goal can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_t + \lambda_1 \mathcal{L}_l + \lambda_2 \mathcal{L}_g \tag{10}$$

where $\mathcal{L}_t$ can be MAE loss (Liu et al., 2018) for multimodal sentiment analysis and contrastive loss (Gabeur et al., 2020) for video retrieval.

4 Experiments

4.1 Datasets

We evaluate the performance of MCC on three challenging tasks, including multimodal sentiment analysis, multimodal speaker traits recognition, and multimodal video retrieval.

CMU-MOSI(Zadeh et al., 2016a): The goal of multimodal sentiment analysis is to identify a speaker's sentiment based on the speaker's display of verbal and nonverbal behaviors. There are a total of 2199 data points (opinion utterances) within CMU-MOSI datasets. The dataset has real-valued sentiment intensity annotations in the range $[-3, +3]$. It is considered a challenging dataset due to speaker diversity (1 video per distinct speaker), topic variations and low-resource setup.

POM(Pérez-Rosas et al., 2013): The dataset contains 963 movie review videos, it is designed for speaker traits recognition based on communicative behavior of a speaker. There are 16 different speaker traits in total. Each video is annotated for various personality and speaker traits.

MSR-VTT(Xu et al., 2016): MSR-VTT is composed of 10k videos. Each video is 10 to 30s long, and is paired with 20 natural sentences describing it. We report results on the train/test splits introduced in (Gabeur et al., 2020) that uses 9000 videos for training and 1000 for testing.

4.2 Experiments for Sentiment Analysis and Speaker Traits Recognition

Data Preprocessing: We extract three modalities for CMU-MOSI and POM, including textual (Pennington et al., 2014), visual (iMotions, 2017), and audio modalities (Degottex et al., 2014).

Experimental Details: Transformer-based multimodal interaction methods have gone mainstream in recent years. With the flexible multi-head attention mechanism, the self-modal interactions and cross-modal interactions can be implemented easily. We implement MCC based on Transformer-based backbone and make comparisons with existing methods. Our Transformer backbone named SC-Transformer (The detailed structure is shown in Appendix Fig. 6) is similar to MulT (Tsai et al., 2019) (Appendix Fig. 5), MulT is a commonly-used baseline that first introduces Transformer structure into the multimodal sequential learning. However, MulT separately processes the intra (self)

Table 1: The experimental results on CMU-MOSI, where we use five metrics, including binary accuracy (BA), F1 score, Pearson Correlation Coefficient (Corr), Multi-class accuracy (MA), and Mean-absolute Error (MAE). For BA, F1, Corr, and MA, higher value is better, as for MAE, lower is better.

Model \ Metric	*BA*	*F1*	*MAE*	*Corr*	*MA*
LMF (Liu et al., 2018)	76.4	75.7	0.912	0.668	32.8
MTGAT (Yang et al., 2021)	81.9	81.7	0.881	0.709	39.1
MulT (Tsai et al., 2019)	83.0	82.8	0.870	0.698	40.0
LMF-MulT (Sahay et al., 2020)	82.8	82.7	0.868	0.705	40.0
HGraph (Lin et al., 2022)	81.7	81.8	0.885	0.702	38.7
AMOA (Li et al., 2022)	82.3	82.1	0.881	0.685	38.4
ConFEDE (Yang et al., 2023)	81.5	81.7	0.883	0.694	39.1
SC-Trans.	83.0	82.8	0.874	0.698	39.5
MCC	**83.6**	**83.5**	**0.863**	**0.717**	**40.7**

Table 2: MCC achieves superior performances over baseline models in POM dataset. MA(5,7) denotes multi-class accuracy for (5,7) classes. All the results of 16 traits are shown in the appendix (Sec. F).

Model \ Trait	*Con*	*Pas*	*Voi*	*Dom*	*Cre*	*Viv*	*Exp*	*Ent*
	MA7	*MA7*	*MA7*	*MA7*	*MA7*	*MA7*	*MA7*	*MA7*
LMF (Tsai et al., 2019)	35.9	35.9	34.8	39.6	34.5	35.9	37.8	36.5
MTGAT (Yang et al., 2021)	35.9	35.5	36.5	39.6	34.5	36.9	40.5	37.9
MulT (Tsai et al., 2019)	34.5	34.5	36.5	38.9	37.4	36.9	37.9	39.4
LMF-MulT (Sahay et al., 2020)	34.5	35.5	36.5	39.6	37.4	36.9	37.8	39.4
HGraph (Lin et al., 2022)	35.9	34.5	36.5	38.9	34.5	36.9	37.9	38.9
AMOA (Li et al., 2022)	35.9	34.5	37.4	38.9	37.0	35.9	37.9	38.9
ConFEDE (Yang et al., 2023)	34.5	35.5	37.4	41.9	34.5	36.9	36.0	37.9
SC-Trans.	34.5	34.5	34.8	39.6	37.0	38.7	37.9	38.9
MCC	**39.4**	**36.9**	**37.4**	**44.3**	**37.9**	**41.4**	**40.9**	**40.4**

and inter (cross) interactions in tandem, which does not fit in with the precondition of MCC (i.e., in parallel). Thus, we simply add a self-attention module in the cross-modal interaction block, making the self and cross attention parallel. The hyperparameters include Adam learning rate 0.001, the structure of projection network (where the hidden size is 40, the size of common space is 16, the number of random features is 64). λ_1 and λ_2 are set to 0.1 and 0.01. The temperature for contrastive learning is set to 0.2. We conduct the experiments on RTX 3080Ti GPUs.

Compared Baselines: We mainly compare MCC with the baseline methods that utilize the same features (i.e. Glove, FACET, COVAREP) for fairness. We reproduce the experimental results that do not be conducted on CMU-MOSI and POM by ourselves.

Experimental Results: The results are shown in Table 1. We can find that MCC consistently gains the best performance across all the baseline methods. We provide a detailed analysis for such observations. MTGAT, which develops graph convolutional networks to capture the multimodal interactions, performs worse than MCC as it pays more attention to the complexity reduction. As for the tensor based multimodal fusion methods, LMF neglects the fine-grained temporal interaction which includes rich structured information for multimodal modeling. Further, MCC outperforms SC-Transformer, LMF-MulT, HGraph, AMOA, ConFEDE with a better overall performance. In general, the best performances of MCC are attribute to the local fine-grained contrastive constraints and the complementary global coarse-grained contrastive constraints which make the attention capture more

Table 3: Ablation study on CMU-MOSI dataset.

Model \ Metric	*BA*	*F1*	*MAE*	*Corr*	*MA*
SC-Trans.	83.0	82.8	0.874	0.698	39.5
w/o. Global	82.6	82.8	0.867	0.705	40.1
w/o. Local	83.0	82.8	0.870	0.704	39.8
w/o. SEA	83.4	83.3	0.865	0.718	40.5
MCC	83.6	83.5	0.863	0.717	40.7

Table 4: Ablation study on POM dataset.

Metric \ Module	SC-Trans.	w/o. Global	w/o. Local	w/o. SEA	Ours
MA (average)	42.0	43.5	42.7	44.2	44.3

Table 5: Complexity of different variants on CMU-MOSI, we only consider the FLOPs of specific modules.

Metric \ Module	Local	Local (SEA)	Global	Global (SEA)
FLOPs	2.3×10^5	3.8×10^4	1.5×10^6	1.0×10^4

accurate[1]. Table 2 shows the experimental results of different methods on speaker traits recognition dataset, POM, where we report the multi-class accuracy of all the traits. A similar observation is found from the table, MCC achieves competitive performances compared with all the baseline methods on most of the traits. Particularly, the performance of MCC increases the average multi-class accuracy from 42.0 to 44.3 compared to the best counterparts. Further, the improvement on POM is larger than CMU-MOSI. At first, we think the reason is the size of dataset. However, when we try the larger datasets like CMU-MOSEI (The results are shown in section I of appendix), the improvement is similar to CMU-MOSI. Thus, we think the improvement of POM is caused by the multi-task training mechanism, since we train 16 traits together, which is different from CMU-MOSI.

Ablation Study: We set some control experiments on CMU-MOSI and POM to verify the effectiveness of MCC and the results are shown in Tables 3, 4, and 5, where "w/o. Global" denotes the model without global contrastive constraints, "w/o. Local" denotes the model without local contrastive constraints, "w/o. SEA" denotes the model without SEA approximation. We could observe that "w/o. SEA" performs similarly to MCC, since the approximation mechanism mainly focus on the reduction of computational complexity. Besides, MCC and "w/o. SEA" perform much better than "w/o. Global" and "w/o. Local", since the proposed global and local contrastive constraints are complementary to each other, only one of them can not lead to large improvement. We evaluate the stability of SEA approximation and the results are shown in the appendix (Sec. D). We could find that, when the number of random features is big enough, MCC achieves competitive performances of MA close to "w/o. SEA". We also provide some visualization results to show the effectiveness of modality alignments in the appendix (Sec. G). We mainly compare the results of SC-Transformer and MCC. Each word in the sentence can more accurately attend to the visual and audio modalities when using contrastive alignment constraints[1].

4.3 Experiments for Multimodal Retrieval

For evaluating the scalability of MCC, we conduct experiments on MSR-VTT with 7 modalities and larger size. Please refer to appendix (Sec. H).

5 Conclusion

In this paper, we propose a generalizable constrained scheme for multimodal sequential learning, which could align attentional sequences from both local and global perspectives. Concretely, the multimodal contexts at the same sequential step are considered as a positive set and the remaining sets are negative. We adopt additional random feature mechanism to approximate the real process, reducing the complexity. We conduct extensive experiments on several traditional tasks, the experimental results reveal the effectiveness of our contributions. In the future, we would focus on how to sample negative groups more effectively and simulate the constrained process without approximation error.

[1] Due to the space limit, we put the attention visualization results in the appendix (Sec. G).

Achknowledgement

This work was supported by National Key R&D Program of China under Grant No.2022ZD0162000.

Limitations and Future Work

We conclude the limitations of MCC as follows: (1) Although we adopt SEA to approximate the contrastive learning process to reduce the complexity, the approximation error always exists. In the future, we would try to devise a more effective unbiased estimation. The goal of unbiased estimation is to make the approximation value close to the real value, avoiding the influence to performance. We think that it need more complicated theoretical derivations to achieve the goal. For examples, borrowing the ideas of effective linear Transformer may be a good direction (Kitaev et al., 2020; Wang et al., 2020). (2) MCC can only be used for sequential input, thus, if the features are not sequential, MCC cannot generalize to. (3) the complexity is related to the sequence length of input, thus, MCC is not compatible with extremely long-sequence input.

In the future, we would focus on how to sample negative groups more effectively and simulate the constrained process without approximation error. (1), Specifically, during the local constraints, we only sample the negative groups where only one time step changes, thus, many negative groups with more changes are wasted. We decide to study more theoretical derivations to make full use of the negative groups. (2), As for the approximation error, it is equivalent to unbiased estimation, The goal of unbiased estimation is to make the approximation value close to the real value, avoiding the influence to performance. We think that it need more complicated theoretical derivations to achieve the goal. For examples, borrowing the ideas of effective linear Transformer may be a good direction (Kitaev et al., 2020; Wang et al., 2020).

Broader Impact

MCC has some real-world applications. First, MCC is plug-and-play and can enhance the alignment of different modalities when conducting multimodal interaction. In fact, many scenarios need the multimodal interaction, like the E-commerce live interaction and short video interaction. Second, MCC can check whether the different modalities are aligned (For example, MCC can be used for detecting fake titles of the videos when the titles not correspond to the videos).

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. 2019. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*.

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/CVF International Conference on Computer Vision*, pages 1728–1738.

Krzysztof Choromanski, Mark Rowland, and Adrian Weller. 2017. The unreasonable effectiveness of structured random orthogonal embeddings. *arXiv preprint arXiv:1703.00864*.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *ICASSP*.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision (ECCV)*, volume 5. Springer.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Paul Hager, Martin J Menten, and Daniel Rueckert. 2023. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23924–23935.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

iMotions. 2017. Facial expression analysis.

Kaixiang Ji, Jiajia Liu, Weixiang Hong, Liheng Zhong, Jian Wang, Jingdong Chen, and Wei Chu. 2022. Cret: Cross-modal retrieval transformer for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 949–959.

Purushottam Kar and Harish Karnick. 2012. Random feature maps for dot product kernels. In *Artificial intelligence and statistics*, pages 583–591. PMLR.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

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

Ziming Li, Yan Zhou, Weibo Zhang, Yaxin Liu, Chuanpeng Yang, Zheng Lian, and Songlin Hu. 2022. Amoa: Global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7136–7146.

Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. 2022. Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7124–7135.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*.

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/CVF International Conference on Computer Vision*, pages 2630–2640.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Andriy Mnih and Yee Whye Teh. 2012. A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426*.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *ACL*.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Saurav Sahay, Eda Okur, Shachi H Kumar, and Lama Nachman. 2020. Low rank fusion based transformers for multimodal sequences. *arXiv preprint arXiv:2007.02038*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321.

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.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. Mtgat: Multimodal temporal graph attention networks for unaligned human multimodal language sequences. In *NAACL*.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *AAAI*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *AAAI*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016a. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016b. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018c. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

A The Detailed Structure of SC-Transformer

To implement MCC, we employ the variant SC-Transformer of MulT. The main difference between these two structure is the usage of self-attention in the cross-modal interaction stage. The comparison is shown in Figs. 5 and 6. In detail, "A→B" denotes multi-head attention mechanism (Vaswani et al., 2017) where B is the query and A denotes key and value.

B The Proof of Proposition 1

Proposition 1: Suppose that we have the group of contexts for the t-th sequential step $c_i^t \in \mathbb{R}^d$ where $i \in [n]$. When we randomly change the sequential step of one specific context, the obtained new group has less relevance than the original positive group. Further, if we treat the score of positive n-vector group consists of the scores of multiple ($\frac{n(n-1)}{2}$) positive two-vector pairs, the scores of negative n-vector groups (only changing one specific context) also contain the scores of all the negative two-vector pairs.

Proof 1: The fact in the first three lines is obvious. Thus, we mainly prove the proposition in the last three lines. We list the negative two-vector pairs obtained by the two methods, simultaneously. Concretely, we select one sequential step t. First, we separately calculate the scores of negative two-vector pairs of $\frac{n(n-1)}{2}$ positive two-vector pairs. We could obtain the following equation:

$$G = \sum_{i=1}^{n}\sum_{j=1,j<i}^{n}\sum_{t'=1,t'\neq t}^{t_a}(A(c_i^t, c_j^{t'}) + A(c_i^{t'}, c_j^t)) \tag{11}$$

where $\{i, j\}$ denotes one modality pair and there are $2(t_a - 1)$ negative two-vector pairs for each $\{i, j\}$ modality pair. Second, we list the negative two-vector pairs in the negative n-vector

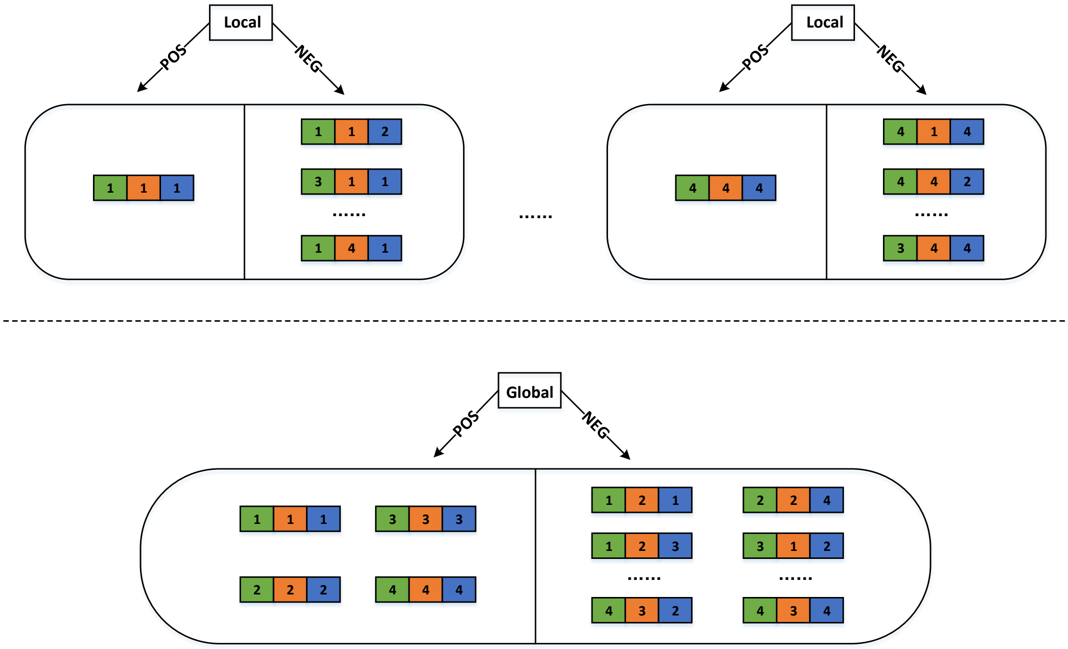


Figure 3: The visual illustration of local and global contrative constraints. The top part denotes local mechanism, where we only change the sequential step of one modality to create negative objects. The bottom part denotes global mechanism, where we divide all the groups into two sets, positive and negative.

groups and we only change the sequential step of one specific context. Suppose we change the sequential step of k-th modality, the relevance scores of new groups are $A(\{c_i^{t'_i}|i \in [n]\}) = \sum_i^n \sum_{j=1,j<i}^n A(c_i^{t'_i}, c_j^{t'_j})$, where $t'_i = t$ for $i \in [n], i \neq k$ and $t'_k \neq t$. During the process, multiple negative two-vector pairs appear and can be expressed as:

$$g' = \sum_{i=1,i\neq k}^{n} \sum_{t'_k=1,\neq t}^{t_a} A(c_i^t, c_k^{t'_k}) \tag{12}$$

We can obtain corresponding negative two-vector pairs by changing the sequential steps of other modalities. The total set can be expressed as:

$$G' = \sum_{k=1}^{n} \sum_{i=1,i\neq k}^{n} \sum_{t'_k=1,\neq t}^{t_a} A(c_i^t, c_k^{t'_k}) \tag{13}$$

We can easily find that sets G and G' are equal. Each element in G can also be found in G', vice versa. Thus, the proposition is proved.

C The Derivations of Eq. 9 of Main Paper

$$\begin{aligned}&\exp(\sum_{i=1}^{n}\sum_{j=1,<i}^{n}\langle v_i, v_j\rangle)\\&=\exp(\frac{\|v_1+...+v_n\|^2}{2})\cdot\prod_{i=1}^{n}\exp(-\frac{\|v_i\|^2}{2})\end{aligned} \tag{14}$$

Next, let $w \in \mathbb{R}^d$. We use the fact that:

$$(2\pi)^{-d/2}\int \exp\left(-\|w-c\|_2^2/2\right)dw = 1 \tag{15}$$

for any $c \in \mathbb{R}^d$ and derive:

$$\begin{aligned}&\exp(\frac{\|v_1+...v_n\|^2}{2}) = (2\pi)^{-d/2}\exp(\frac{\|v_1+...v_n\|^2}{2})\cdot\\&\cdot\int\exp(\frac{-\|w-(v_1+...v_n)\|^2}{2})dw\\&=(2\pi)^{-d/2}\int\exp(-\frac{\|w\|^2}{2}+w^\top(v_1+..v_n)\\&-\frac{\|v_1+..v_n\|^2}{2}+\frac{\|v_1+...v_n\|^2}{2})dw\\&=(2\pi)^{-d/2}\int\exp(-\frac{\|w\|^2}{2}+w^\top(v_1+...v_n))dw\\&=(2\pi)^{-d/2}\int\exp(\frac{-\|w\|^2}{2})\cdot\prod_{i=1}^{n}\exp(w^\top v_i)dw\\&=\mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I}_d)}[\prod_{i=1}^{n}\exp(w^\top v_1)]\end{aligned} \tag{16}$$

That completes the proof. We then provide the estimation error of the approximation. Suppose that the number of random features is H,

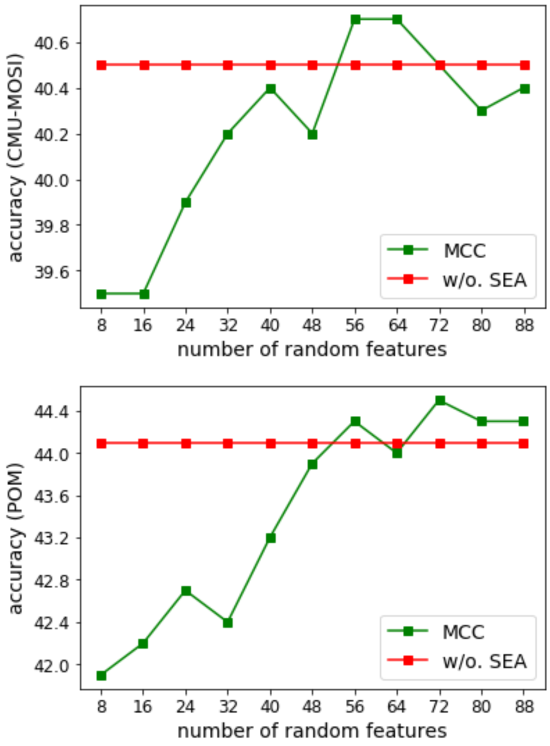


Figure 4: The evaluation of random features.

$$\begin{aligned}&\exp(\sum_{i=1}^{n}\sum_{j=1,<i}^{n}\langle v_i, v_j\rangle)\\ &= \exp\left(-\frac{\|v_1\|^2+...+\|v_n\|^2}{2}\right)\cdot\\ &\cdot\,\mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{1}_d)}\left[\exp\left(w^\top\mathbf{z}\right)\right]\end{aligned} \tag{17}$$

where $\boldsymbol{z} = v_1 + v_2 + ... + v_n$, based on the fact $\mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\exp\left(w^\top\mathbf{z}\right)\right] = \exp\left(\frac{\|\mathbf{z}\|^2}{2}\right)$, then we can obtain:

$$\begin{aligned}&\mathrm{MSE}(\exp_H(\sum_{i=1}^{n}\sum_{j=1,<i}^{n}\langle v_i, v_j\rangle))\\ &= \frac{1}{H}\exp\left(-\beta\right)\mathrm{Var}\left(\exp\left(w^\top\mathbf{z}\right)\right)\\ &=\frac{1}{H}\exp(-\beta)(\mathbb{E}[\exp(2w^\top\mathbf{z})]-(\mathbb{E}[\exp(w^\top\mathbf{z})])^2)\\ &= \frac{1}{H}\exp\left(-\beta\right)\left(\exp\left(2\|\mathbf{z}\|^2\right)-\exp\left(\mathbf{z}^4\right)\right)\\ &= \frac{1}{H}\exp\left(-\beta\right)\exp\left(\|\mathbf{z}\|^2\right)\left(\exp\left(\|\mathbf{z}\|^2\right)-1\right)\\ &= \frac{1}{H}\gamma\exp(\sum_{i=1}^{n}\sum_{j=1,<i}^{n}\langle v_i, v_j\rangle)^2\left(1-\frac{1}{\gamma}\right)\\ &\beta = \|v_1\|^2+\|v_2\|^2...+\|v_n\|^2\\ &\gamma = \exp\left(\|\mathbf{z}\|^2\right)\end{aligned} \tag{18}$$

where H denotes the number of random features. It is obvious that improving H can help reduce the approximation error.

To further reduce the variance of the estimator, we entangle different random weights $w_1, \ldots, w_H$ to be exactly orthogonal. This can be done while maintaining unbiasedness whenever isotropic distributions $\mathcal{N}(0, \mathbf{I}_d)$ are used by standard Gram-Schmidt renormalization procedure (Choromanski et al., 2017). ORFs is a well-known method and can be applied to reduce the variance of softmax/Gaussian kernel estimators for any dimensionality d rather than just asymptotically for large enough d and leads to the first exponentially small bounds on large deviations probabilities that are strictly smaller than for non-orthogonal methods. The ORF mechanism requires $H \le d$, if $H > d$, ORFs still can be used locally within each $d \times d$ block.

D Ablation Study of the Number of Random Features

According to Eq. 18, the approximation error is related to the sampling number, increasing the sampling number can reduce the error.

The ablation results are shown in Fig. 4. The red line denotes the method without SEA approximation, the complexity is $O(10^5)$ as shown in Table 5, while MCC with much lower complexity $O(10^4)$ (smaller sampling number) can achieve similar or better results than the model without SEA, which also means the approximation error is small.

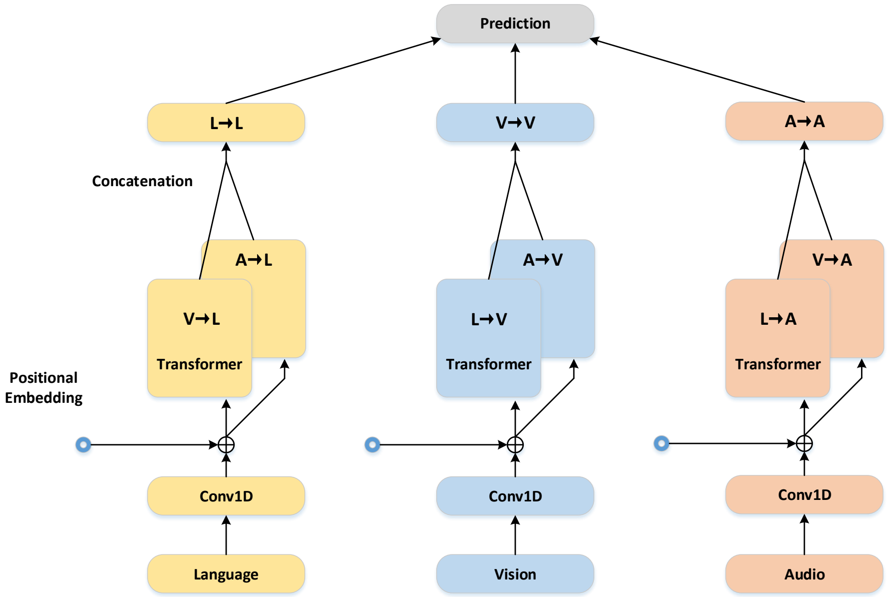


Figure 5: The structures of MulT (Tsai et al., 2019). "A→B" denotes multi-head attention mechanism (Vaswani et al., 2017) where B is the query and A denotes key and value.

E Illustration of Complexity Analysis

We detailedly analyze the alignment complexity before and after the approximation. The total complexity of the global contrastive alignments is $O(n^2(t_a)^n)$, the square complexity $O(n^2)$ arises from the combinational addition operations and the exponential complexity $O((t_a)^n)$ arises from the large number of context sets. We argue that the SEA approximation can reduce the square complexity to a linear level $O(n)$. Specifically, the relevance score of a set can be calculated with continuous multiplication operation according to the Eq. 9 (main paper). As for the exponential part, we calculate the sum of all the relevance scores like $\sum_i \sum_j \sum_k a_i b_j c_k = (\sum_i a_i)(\sum_j b_j)(\sum_k c_k)$. Therefore, the complexity with approximation becomes $O(n^2 t_a)$.

According to Eq. 9 (main paper), we can obtain the following equation:

$$\exp(A(\{c_i^t | i \in [n]\})) = \text{Mean}(\prod_{i=1}^{n} (c_i^t)') \quad (19)$$

where $(c_i^t)'$ denotes the transformation of c_i^t with Eq. 9 (main paper) and $Mean()$ denotes the mean operation for the elements in the vector. Therefore, the sum of the scores of all the n-vector group is as follows:

$$\begin{aligned}
&\sum_{t'_1=1}^{t_a}\sum_{t'_2=1}^{t_a}\dots\sum_{t'_n=1}^{t_a}\exp(A(\{c_i^{t'_i}|i\in[n]\}))\\
&=\sum_{t'_1=1}^{t_a}\sum_{t'_2=1}^{t_a}\dots\sum_{t'_n=1}^{t_a}\text{Mean}(\prod_{i=1}^{n}(c_i^{t'_i})')\\
&=\text{Mean}(\sum_{t'_1=1}^{t_a}\sum_{t'_2=1}^{t_a}\dots\sum_{t'_n=1}^{t_a}\prod_{i=1}^{n}(c_i^{t'_i})')\\
&=\text{Mean}(\prod_{i=1}^{n}(\sum_{t'_i}^{t_a}(c_i^{t'_i})'))
\end{aligned} \quad (20)$$

Therefore, we change the complexity from $O(n^2(t_a)^n)$ to $O(n^2 t_a)$.

F The Complete Results of POM

We shown the results of all the 16 traits in Table 6.

G Attention Visualization

We also provide the visualization results in Fig. 7 and Fig. 8, which show that MCC can help the multimodal interaction occurs between the more related segments of different modalities. We take the word "fighting" as an example, MCC can accurately locate the word in the corresponding positions of visual and audio modalities, while SC-Transformer can not.

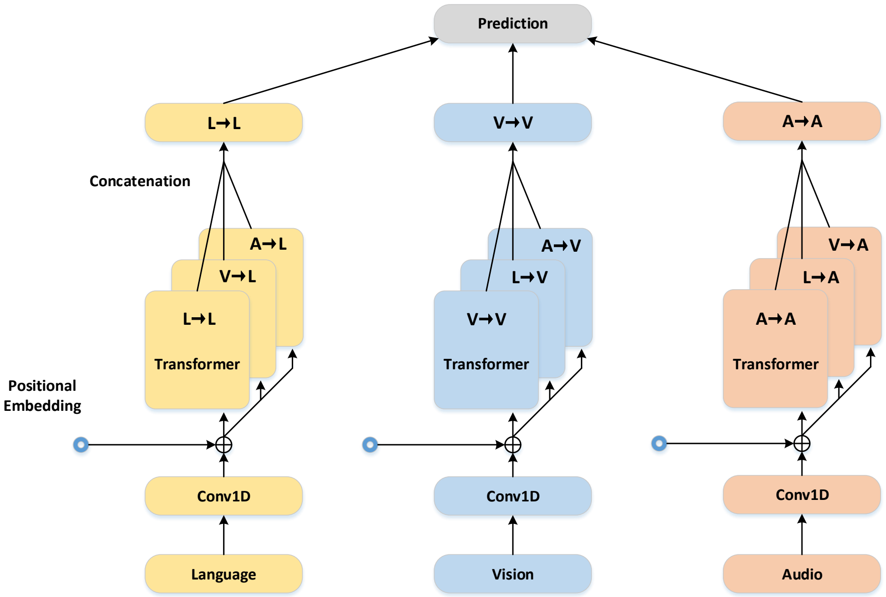


Figure 6: The structures of SC-Transformer. "A→B" denotes multi-head attention mechanism (Vaswani et al., 2017) where B is the query and A denotes key and value.

H Experiments for Multimodal Video Retrieval

Data Preprocessing: we follow (Gabeur et al., 2020)and use multiple pre-trained models for extracting features. Concretely, we utilize following 7 experts: **Motion** embeddings are extracted from S3D (Xie et al., 2018) trained on the kinetics dataset. **Scene** embeddings are extracted with DenseNet-161 (Huang et al., 2017) trained on the Place365 dataset (Zhou et al., 2017). **OCR** embeddings are encoded with a word2vec embedding. **Audio** embeddings are obtained with a VGGish model, trained on the YouTube-8m dataset. **Speech** features are extracted using the Google Cloud speech API, to extract word tokens from the audio stream, which are then encoded via pre-trained word2vec (Mikolov et al., 2013a) embeddings. **Face** features are extracted by ResNet-50 (He et al., 2016) trained on the VGGFace2 dataset. **Appearance** features are extracted by SENet-154 (Hu et al., 2018) trained on ImageNet.

Experimental Details: We implement MCC based on the Transformer based backbone and make comparisons with existing methods. We utilize MMT as backbone. We implement MCC by normalizing the attention weights along the sequential steps of corresponding modalities in parallel. The hyperparameters of MCC include Adam learning rate 5×10^{-5}, which we decay by a multiplicative factor 0.95 every 1000 optimization steps, the structure of projection network (where the hidden size is 512, the size of common space is 64, the number of random features is 512). λ_1 and λ_2 are set to 0.1 and 0.01. The temperature for contrastive learning is set to 0.2. We conduct all the experiments on RTX 3080Ti GPUs (10GB).

Compared Baselines: For fairness, we mainly compare MCC with the baseline methods CE, Dual Enc, FIT, CLIPBERT, CERT, MMT that do not utilize large-scale pre-training with HowTo100M (Miech et al., 2019) dataset.

Experimental Results: We report the evaluation results of MCC and the competing text-video retrieval methods on MSR-VTT (Table 7). Our MCC performs constantly better than other baselines (Dual Enc, FIT, CLIPBERT, CERT, ConFEDE, MMT). Specifically, the R@5 score of MCC can reach 56.4% and 57.8% of text-to-video and video-to-text tasks, making the relative improvement over the best competitor by 1.2% and 1.6%. As expected, MCC utilizing both global and local contrastive constraints exhibits better performance than that only using the indirect alignments of attention mechanism.

I More Experiments on Other Datasets

I.1 CMU-MOSEI

The CMU-MOSEI dataset (Zadeh et al., 2018c) is the next generation of CMU-MOSI. Compared with CMU-MOSI, it has much larger training sam-

Table 6: MCC achieves superior performances over baseline models in POM dataset (multimodal personality traits recognition). MA(5,7) denotes multi-class accuracy for (5,7) classes.

Model \ Trait	*Con*	*Pas*	*Voi*	*Dom*	*Cre*	*Viv*	*Exp*	*Ent*
	MA7	*MA7*	*MA7*	*MA7*	*MA7*	*MA7*	*MA7*	*MA7*
LMF (Tsai et al., 2019)	35.9	35.9	34.8	39.6	34.5	35.9	37.8	36.5
MTGAT (Yang et al., 2021)	35.9	35.5	36.5	39.6	34.5	36.9	40.5	37.9
MulT (Tsai et al., 2019)	34.5	34.5	36.5	38.9	37.4	36.9	37.9	39.4
LMF-MulT (Sahay et al., 2020)	34.5	35.5	36.5	39.6	37.4	36.9	37.8	39.4
HGraph (Lin et al., 2022)	35.9	34.5	36.5	38.9	34.5	36.9	37.9	38.9
AMOA (Li et al., 2022)	35.9	34.5	37.4	38.9	37.0	35.9	37.9	38.9
ConFEDE (Yang et al., 2023)	34.5	35.5	37.4	41.9	34.5	36.9	36.0	37.9
SC-Trans.	34.5	34.5	34.8	39.6	37.0	38.7	37.9	38.9
MCC	**39.4**	**36.9**	**37.4**	**44.3**	**37.9**	**41.4**	**40.9**	**40.4**

Model \ Trait	*Res*	*Tru*	*Rel*	*Out*	*Tho*	*Ner*	*Per*	*Hum*
	MA5	*MA5*	*MA5*	*MA5*	*MA5*	*MA5*	*MA7*	*MA5*
LMF (Tsai et al., 2019)	35.5	54.2	53.2	44.8	42.7	43.5	34.9	45.8
MTGAT (Yang et al., 2021)	36.9	55.7	54.2	44.8	46.0	44.8	37.8	43.5
MulT (Tsai et al., 2019)	41.4	60.6	54.2	43.3	**49.3**	46.3	33.5	43.3
LMF-MulT (Sahay et al., 2020)	41.4	57.1	54.2	44.8	47.3	46.3	37.8	43.5
HGraph (Lin et al., 2022)	38.4	55.7	54.2	46.8	47.3	44.8	34.9	45.8
AMOA (Li et al., 2022)	39.6	60.6	53.2	46.8	46.5	46.3	37.8	45.8
ConFEDE (Yang et al., 2023)	38.4	57.1	53.2	46.8	47.3	47.8	34.0	47.3
SC-Trans.	39.6	59.5	**55.2**	47.4	46.5	47.0	34.9	44.8
MCC	**41.9**	**61.6**	51.2	**50.7**	45.8	**48.3**	**46.3**	**49.8**

ples and more variations in speakers and video topics. Specifically, CMU-MOSEI contains 23453 manually annotated utterance-level video segments from 1000 distinct speakers and 250 different topics.

The experimental details of CMU-MOSEI are similar to CMU-MOSI. We obtain the results in Table 8. MCC still outperform all the compared methods with a better overall performance.

I.2 LSMDC

LSMDC contains 118081 short video clips (about $45s$) extracted from 202 movies. Each clip is annotated with a caption, extracted from either the movie script or the audio description. The test set is composed of 1000 videos, from movies not presented in the training set.

The experimental details of LSMDC are similar to MSR-VTT. We obtain the results in Table 9. MCC still outperform all the compared methods with a better overall performance.

Table 7: Retrieval performances on the MSR-VTT, where we employ R@K and MdR as metrics.

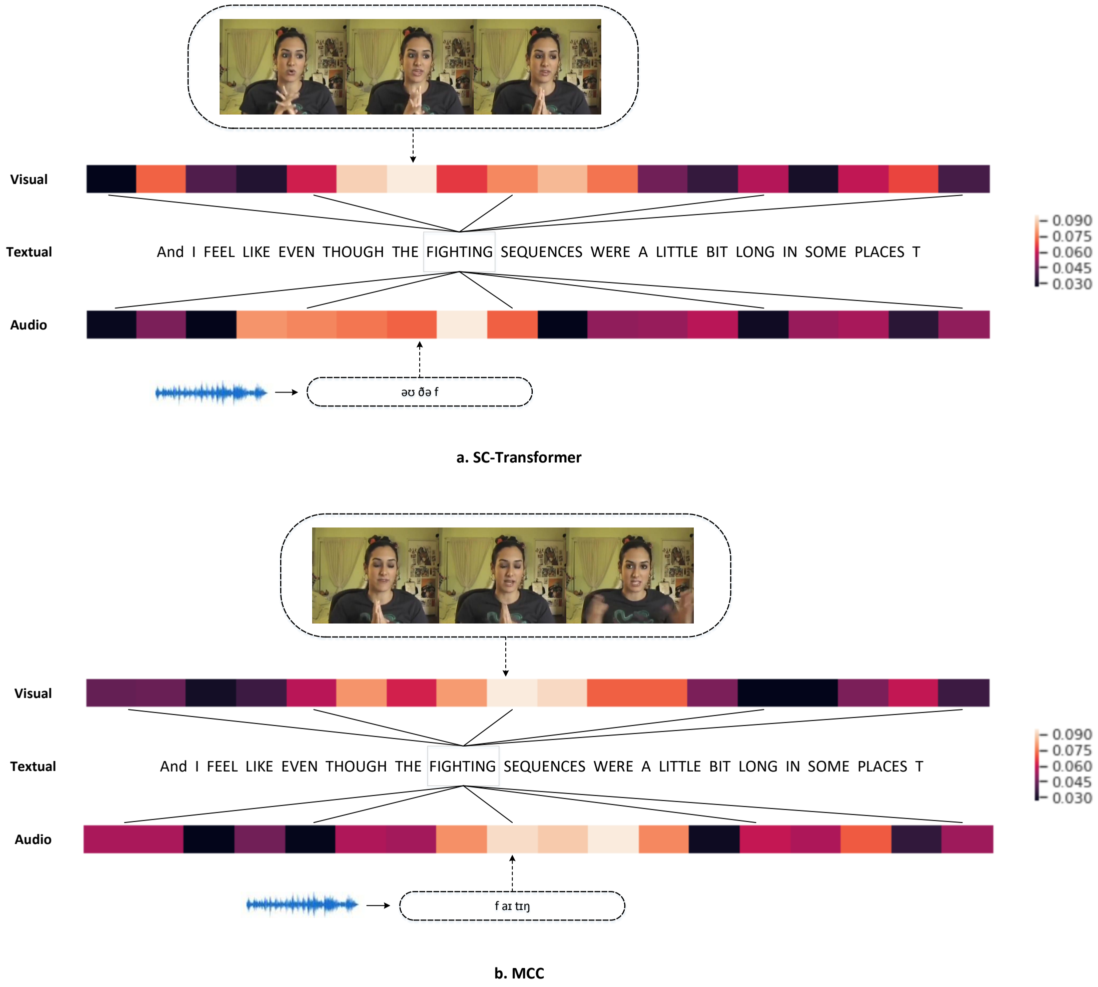
Model \Metric	*Text ⟶ Video*				*Video ⟶ Text*			
	R@1↑	*R@5*↑	*R@10*↑	*MdR*↓	*R@1*↑	*R@5*↑	*R@10*↑	*MdR*↓
Dual Enc (Dong et al., 2021)	23.0	50.6	62.5	5	25.1	52.1	64.6	5
FIT (Bain et al., 2021)	15.2	-	54.4	9	-	-	-	-
CLIPBERT (Lei et al., 2021)	22.0	46.8	59.9	6	-	-	-	-
MMT (Gabeur et al., 2020)	24.6	54.0	67.1	4	24.4	56.0	67.8	4
CRET (Ji et al., 2022)	23.9	50.8	63.4	5	-	-	-	-
ConFEDE (Yang et al., 2023)	24.5	55.2	67.5	4	24.5	56.2	67.5	4
MCC	**25.4**	**56.4**	**69.1**	**3**	**25.5**	**57.8**	**68.3**	**3**

Table 8: The experimental results on CMU-MOSEI, where we use five metrics, including binary accuracy (BA), F1 score, Pearson Correlation Coefficient (Corr), Multi-class accuracy (MA), and Mean-absolute Error (MAE). For BA, F1, Corr, and MA, higher value is better, as for MAE, lower is better.

Model \ Metric	*BA*	*F1*	*MAE*	*Corr*	*MA*
LMF (Liu et al., 2018)	80.5	80.3	0.632	0.668	48.2
MTGAT (Yang et al., 2021)	81.8	81.6	0.609	0.689	50.8
MulT (Tsai et al., 2019)	82.5	82.3	0.580	0.703	51.8
HGraph (Lin et al., 2022)	81.7	81.8	0.885	0.702	38.7
ConFEDE (Yang et al., 2023)	81.4	81.3	0.604	0.692	50.3
SC-Trans.	82.2	82.1	0.585	0.705	52.0
MCC	**83.2**	**83.1**	**0.570**	**0.714**	**52.9**

Table 9: Retrieval performances on the LSMDC, where we employ R@K and MdR as metrics.

Model \Metric	*Text ⟶ Video*				*Video ⟶ Text*			
	R@1↑	*R@5*↑	*R@10*↑	*MdR*↓	*R@1*↑	*R@5*↑	*R@10*↑	*MdR*↓
MMT (Gabeur et al., 2020)	13.2	29.2	38.8	21	12.1	29.3	37.9	22.5
CRET (Ji et al., 2022)	10.0	24.9	33.4	34	-	-	-	-
ConFEDE (Yang et al., 2023)	13.5	29.6	39.2	21	12.3	29.7	38.5	22
MCC	**14.4**	**30.8**	**40.4**	**20**	**12.9**	**30.8**	**39.7**	**21**

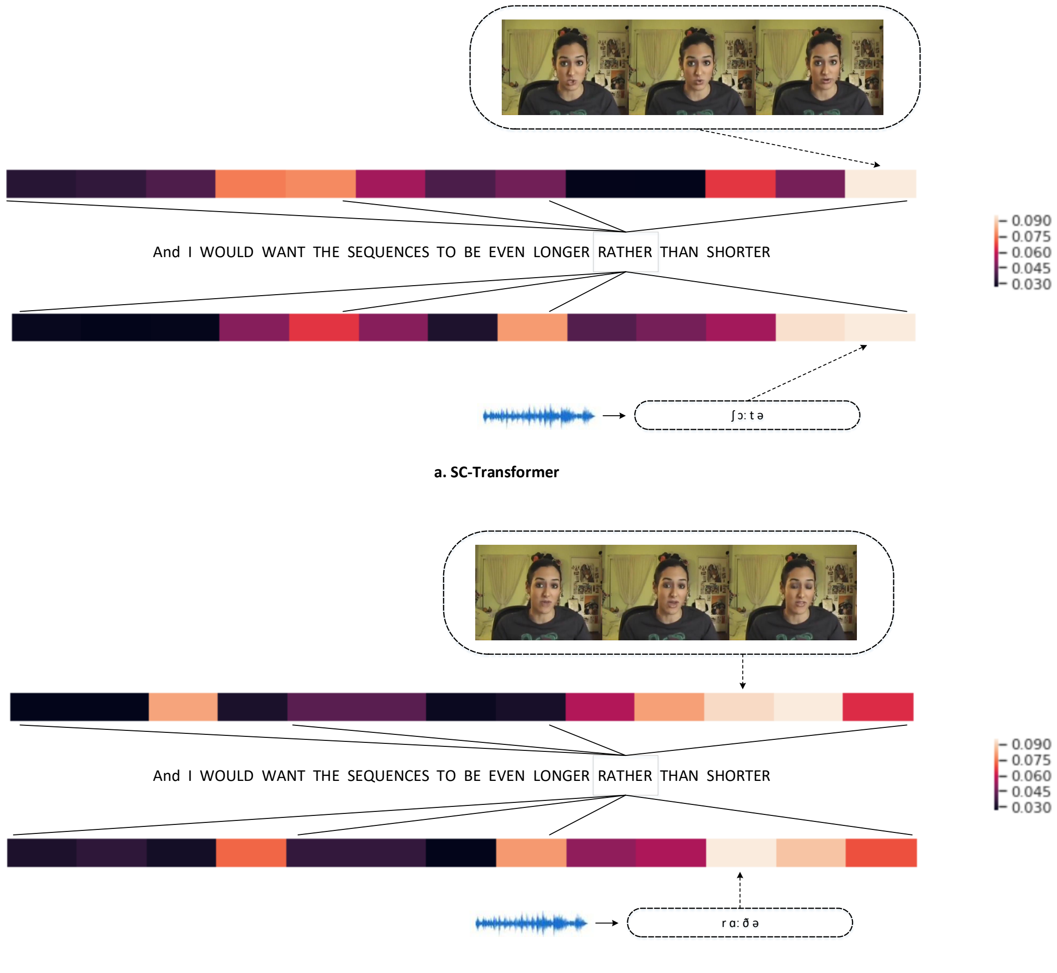

Visual
Textual
And I FEEL LIKE EVEN THOUGH THE FIGHTING SEQUENCES WERE A LITTLE BIT LONG IN SOME PLACES T
Audio
əʊ ðə f
0.090
0.075
0.060
0.045
0.030

a. SC-Transformer

Visual
Textual
And I FEEL LIKE EVEN THOUGH THE FIGHTING SEQUENCES WERE A LITTLE BIT LONG IN SOME PLACES T
Audio
f aɪ tɪŋ
0.090
0.075
0.060
0.045
0.030

b. MCC

Figure 7: A visualization example of multimodal interaction, where we provide the attention weights of cross-modal interaction. For the specfic word "fighting", MCC can localize a more accurate phonetic alphabet and corresponding mouth shapes.



a. SC-Transformer

b. MCC

Figure 8: Another example.