Hierarchical Crossmodal Transformer with Modality Gating for Incongruity-Aware Sentiment and Emotion Recognition

Anonymous ACL submission

Abstract

Fusing multiple modalities for affective computing tasks has proven effective for performance improvement. However, how multimodal fusion works and how to use it are not well understood. Also, multimodal fusion has brought difficulties in real-world use due to the large model sizes. In this work, on sentiment and emotion recognition, we first analyze how the salient affective information in one modality can be affected by the other in crossmodal attention. We find that inter-modal incongruity exists at the latent level due to crossmodal attention. Based on this finding, we propose a lightweight model via Hierarchical Crossmodal Transformer with Modality Gating (HCT-MG), which determines a primary modality according to its contribution to the target task and then hierarchically incorporates auxiliary modalities to alleviate inter-modal incongruity and reduce information redundancy. The experimental evaluation on three benchmark datasets: CMU-MOSI, CMU-MOSEI, and IEMOCAP verifies the efficacy of our approach, showing that it: 1) outperforms major prior work by achieving competitive results and recognizing hard samples; 2) mitigates the inter-modal incongruity at the latent level when modalities have mismatched affective tendencies; 3) reduces model size to less than 1M parameters while significantly outperforming existing models of similar sizes.

1 Introduction

As emotions are expressed in complex ways (e.g., face, voice, and language) in human communication, multimodal fusion has become a hot topic in the past decade. Previous studies have proved that by taking advantage of complementary information from multiple modalities, emotion recognition can be more robust and accurate (Xu et al., 2018; Li et al., 2022a). However, several major issues remain unsolved, impeding the true progress of Multimodal Sentiment and Emotion Recognition (MSER). First, multimodal signals often show an unaligned nature, bringing about the asynchrony problem (Tsai et al., 2019). For example, the visual signal usually precedes the audio by around 120ms when people express emotion (Grant and Greenberg, 2001). Second, different modalities may have different or even opposite affective tendencies, which makes emotions difficult to recognize. For example, people sometimes say positive content with a negative voice (e.g., sarcasm) or negative content with a smile (e.g., to be polite).

Prior work has proposed many approaches to tackle these issues. For example, Tsai et al. (2019) introduced the Multimodal Transformer (MulT) model to learn a pair-wise latent alignment with the Transformer structure, which directly attends to low-level features in multiple modalities to solve the asynchrony problem. Wu et al. (2021) proposed an incongruity-aware attention network that focuses on the word-level incongruity between modalities by assigning larger weights to words with incongruent modalities. Nevertheless, to capture as much information as possible for better performance, recent models usually repeatedly fuse certain or all modalities (Liang et al., 2018), resulting in not only redundant information but also large model sizes that hinder their real-world use.

To address this problem, in this paper we propose the Hierarchical Crossmodal Transformer with Modality Gating (HCT-MG), a lightweight multimodal fusion model that can alleviate intermodal incongruity, reduce information redundancy, and learn representations from unaligned modalities at the same time. Specifically, HCT-MG dynamically determines the primary modality based on its contribution to the target task and then hierarchically fuses auxiliary modalities via crossmodal Transformer to obtain the most useful but least amount of information without modality alignment. Before the model implementation, we perform a *feasibility analysis* of the crossmodal Transformer (specifically, its attention mechanism) in multimodal fusion (i.e., how the salient affective information in one modality is affected by the other at the latent level) for the rationality of our model. We also propose HCT, which is built upon empirical knowledge of the relationship among audio, vision, and text modalities without the modality gating for a *comparison analysis*.

The *feasibility analysis* demonstrates that crossmodal attention functions by highlighting the salient affective information in one modality with the help of the other one. However, when modalities have mismatched affective tendencies, crossmodal attention may malfunction by leaving intermodal incongruity at the latent level. The *comparison analysis* confirms the rationality of manually selecting text as the primary modality, which completes the literature of trimodal studies. The experimental evaluations on CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018b), and IEMOCAP (Busso et al., 2008) show that our approach achieves competitive results and alleviates the inter-modal incongruity with a small model size. (Code available[1])

2 Related Work

Among previous approaches, early fusion and late fusion are the most widely used for MSER. However, due to the strict constraint on time synchrony, early fusion does not work well if the input features of multiple modalities differ in their temporal characteristics (Li et al., 2020). On the other hand, since different modalities have been confirmed to be complementary to each other (Chuang and Wu, 2004), the relatedness among them is ignored by late fusion. To this end, tensor fusion, which is performed at the latent level, has become mainstream. For example, Zadeh et al. (2017) introduced a Tensor Fusion Network, that learns both intra- and inter-modality dynamics end-to-end.

Furthermore, with the success of the cross-attention mechanism (Lu et al., 2019), which exchanges key-value pairs in self-attention, a major trend using cross-attention for multimodal fusion has emerged and is usually referred to as *crossmodal attention*. Tsai et al. (2019) proposed a crossmodal attention-based Transformer to provide tensor-level crossmodal adaptation that fuses multimodal information by directly attending to features in other modalities. Zadeh et al. (2019) developed a self-attention- and cross-attention-based Transformer to extract intra-modal and inter-modal emotional information, respectively. Li et al. (2022a) used crossmodal attention with a hierarchical structure to capture lexical features from different textual aspects for speech emotion recognition.

Despite these advances, some issues still remain challenging in multimodal fusion. First of all, different modalities may show mismatched affective tendencies, resulting in inter-modal incongruity – a general problem for MSER tasks. However, the majority of this topic is based on high-level comparison analysis between modalities, such as a person expressing praise while rolling his/her eyes (Wu et al., 2021). There is no evidence that such intermodal incongruity can be tackled at the latent level by crossmodal attention. What's more, to improve the performance of MSER tasks, certain modalities are usually fused repeatedly. Such an operation would bring information redundancy to the model and result in large model sizes, which hinder the real-world use of MSER. With these challenges in mind, we conduct a novel analysis on how crossmodal attention functions or fails, and propose a lightweight yet efficient model.

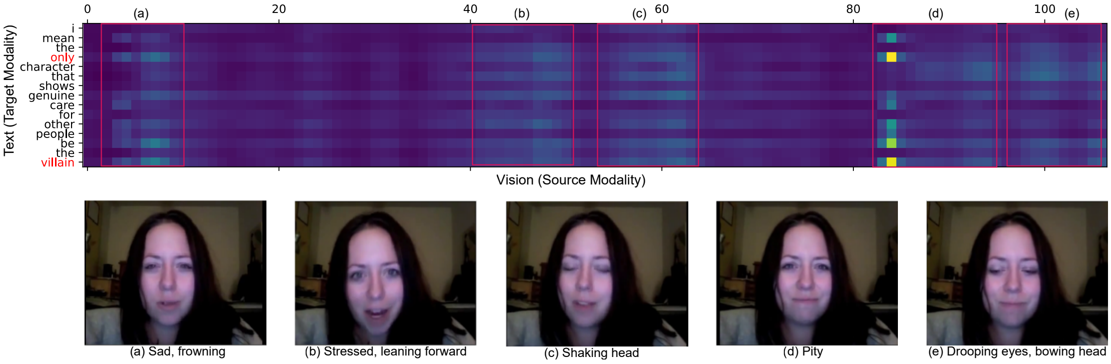
3 Analysis of Crossmodal Attention

Models utilizing data from different modalities usually outperform unimodal ones as more information is aggregated. Prior work has proved that learning with multiple modalities achieves a smaller population risk than only using a subset of modalities. The main reason is that the former, using multimodal data, has a more accurate estimate of the latent space representation (Huang et al., 2021). However, there is no guarantee that learning with multimodal data is always better than unimodal. For example, Huang et al. (2021) found that combining multiple modalities (text, audio, and video) underperforms the unimodal when the number of sample sizes is relatively small. Besides, Rajan et al. (2022) compared a self-attention and a crossmodal attention model for emotion recognition, showing no clear difference between the results of the two models.

As no evidence has been presented as to whether crossmodal attention works and why, we implement an analysis on the latent level to investigate how multimodal information interacts with each other and how inter-modal incongruity occurs. We conduct three experiments on CMU-MOSEI:

Exp 1. Investigating how source modality en-

[1] We will release code on paper acceptance.



(a) Sad, frowning

(b) Stressed, leaning forward

(c) Shaking head

(d) Pity

(e) Drooping eyes, bowing head

Figure 1: Exp 1. Heatmap of highlighted hidden states using crossmodal attention on vision (source) and text (target) modalities.

Crossmodal Attention: but this movie kind of came off like somebody really sad travel video, i don't know

Text Self-attention: but this movie kind of came off like somebody really sad travel video, i don't know

Crossmodal Attention: and umm first of all, i, i actually was never a huge fan of the xx xxx tv series

Text Self-attention: and umm first of all, i, i actually was never a huge fan of the xx xxx tv series

Crossmodal Attention: i really love the melodramatic over the top crazy action sort of thing going on

Text Self-attention: i really love the melodramatic over the top crazy action sort of thing going on

Figure 2: Exp 2. Heatmap of highlighted words by using self-attention with and without crossmodal attention.

Crossmodal Attention ($V \rightarrow T$): i love him as an actor based on his looks, as an actural talented actor, no

Crossmodal Attention ($A \rightarrow T$): i love him as an actor based on his looks, as an actural talented actor, no

Figure 3: Exp 3. Heatmap of highlighted words using different modalities (V or A) in crossmodal attention.

hances target modality via crossmodal attention. We use the example of $V \rightarrow T$ (text attended by vision). *Next, we hope to see how the combination of two modalities affects the third collectively.*

Exp 2. Investigating how the salient parts of the target modality are represented by self-attention with and without the combination of source modalities. We use the example of $(A + V) \rightarrow T$ (text attended by cross-attention-fused audio-vision). *Further, we would like to know how different source modalities affect the target individually.*

Exp 3. Investigating how the salient parts of the target modality are represented by crossmodal attention when using different source modalities. We use the examples of $V \rightarrow T$ (text attended by vision) and $A \rightarrow T$ (text attended by audio).

The experimental setup is shown in Table 1, and the visualization is shown in Figure 1, 2, and 3.

Table 1: Experimental setup for crossmodal attention analysis. T for Text, A for Audio, and V for Vision.

Exp.	Target modality	Source modality	Crossmodal	Self-attention
1	T	V	$V \rightarrow T$	/
2	T	$A + V$	$(A + V) \rightarrow T$	T
3	T	A or V	$A \rightarrow T, V \rightarrow T$	/

In Figure 1, X-axis is the video frames and Y-axis is the text words. The salient emotional information captured by crossmodal attention is highlighted in the red box. It can be noticed that the highlighted parts are due to obvious facial or behavior changes of the character in the video, such as frowning or shaking head. The crossmodal attention successfully highlights the meaningful words that show pity emotion (e.g., "only", "vallain").

In Figure 2, it can be noted that when fused with the combination of source modalities, text focuses more on the words related to emotional information with less noise from other words. For example, when with crossmodal attention, the word "sad" is the most salient in the first sentence, yet much less focused when without. The same is true for the word "never" in the second sentence and the words "love" and "melodramatic" in the third sentence.

In Figure 3, we can see that when fused with different individual source modalities, the target modality (text) can be enhanced with disparate affective tendencies. When using vision as the source modality, the words "love" and "talented" are the most highlighted, representing a positive meaning.

When using audio, however, “no” is the most focused word, showing a negative tendency. This phenomenon demonstrates that different modalities may contain mismatched affective tendencies. The existence of inter-modal incongruity has been found by high-level inter-modal comparison (Desai et al., 2022) and sentiment analysis (Li et al., 2019). Our finding demonstrates that the incongruity also exists at the latent level when using crossmodal attention, resulting in salient affective information in one modality being distorted by the other.

Based on the above findings, we can find that crossmodal attention does help multimodal fusion by aligning two modalities to highlight the salient affective information in the target modality with complementary information from the source. According to the attention mechanism (Vaswani et al., 2017), this process can be described as mapping the Query (from the target) to the Key (from the source) and obtain scores for the Value (from the source). However, such a process could malfunction if the modalities have mismatched affective tendencies, which leaves the inter-modal incongruity difficult to resolve at the latent level.

4 Proposed Approach

In order to exert the advantages of crossmodal attention while solving the above problems, we propose a new multimodal fusion approach: the Hierarchical Crossmodal Transformer (HCT). Unlike previous studies that equally treated all modalities or fused them in every step even the major modality and weights are determined (Rahman et al., 2020), HCT selects a primary modality leaving it for fusion in the final step, while fusing the auxiliary modalities first. Specifically, audio and vision are attended to each other first, and then attended to text. The selection is operated both empirically (by ourselves) and automatically (by the model), bringing two architectures with minor difference.

4.1 Empirical Primary Modality Selection – HCT Model

We manually select text as the primary modality and audio and vision as the auxiliary modalities, inspired by empirical knowledge that text is relatively independent from the other modalities but significantly contributes to affective polarity (Lindquist et al., 2015). Specifically:

1) There is a clear temporal pattern when people express emotions via vision and audio modalities: visual signals usually precede audio by around 120ms (Grant and Greenberg, 2001). 2) People can behave quite differently from what they say in spoken dialogs. For example, positive behaviors sometimes come along with a negative sentence to ease the embarrassment (Li et al., 2019), and a positive sentence can be said in a negative way to express sarcasm (Castro et al., 2019). 3) In MSER applications, a misrecognition of emotion at the level of sentiment polarity would lead to a fatal error (imagine that the system responds "Good to hear that!" in a happy voice when the user in fact feels sad). On the other hand, misclassifying an emotion as another that has the same polarity may well be tolerable (Tokuhisa et al., 2008; Li, 2018). 4) Sentiment largely depends on textual information (Lindquist et al., 2015). Using text as the major modality in fine-tuning and shifting the language-only position of a word to the new position in light of audio-visual information allows the language models (BERT, XLNet) to better yield sentiment scores (Rahman et al., 2020). 5) Modality refers to the way in which something is expressed or perceived. Unlike audio and vision, which are raw (low-level) modalities closest to sensors, text is a relatively abstract and high-level modality that is farther from sensors (Baltrušaitis et al., 2018). According to the nature of the human brain’s hierarchical perceptual processing, low-level information is processed first, followed by high-level information (Peelle et al., 2010). Tian et al. (2016) and Li et al. (2022a) demonstrated that using a hierarchical model to process low-level features and fuse high-level ones step-by-step can yield better representations for emotion recognition.

Inspired by the above studies, we fuse audio and vision modalities to learn their relatedness first, keeping the text modality intact and fusing it in a later step to obtain better representations. This hierarchical process is expected to mitigate the inter-modal incongruity as the fine-grained emotions will not shift much from their sentiment polarities.

The architecture is shown as Figure 4. HCT is constructed based on three modalities: Text (T), Audio (A), and Vision (V), and mainly consists of three components: feature projection, crossmodal Transformer, and weighted concatenation.

Feature Projection. The input features are first fed into 1D Convolutional (Conv1D) networks to integrate local contexts and project the features into the same hidden dimension. Then the fea-

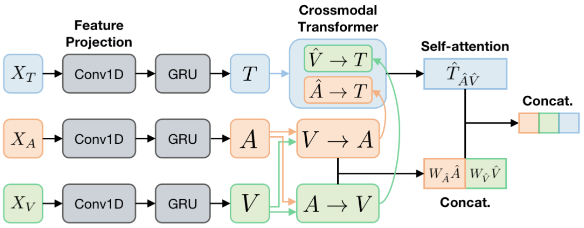


Figure 4: Architecture of HCT.

tures are passed to the Gated Recurrent Unit (GRU) networks, which encode global contexts by updating their hidden states recurrently and model the sequential structure of the extracted features accordingly. We use GRU as it performs similarly to long short-term memory but is computationally cheaper, which is crucial for real-life use.

Crossmodal Transformer. As a variant of self-attention, cross-attention (Lu et al., 2019) transforms the signals from the source modality into a different set of Key-Value pairs to interact with the target modality, which has proven useful in various domains (Zhang et al., 2022; Rashed et al., 2022). The crossmodal Transformer used here is the same as MulT (Tsai et al., 2019), which is a deep stacking of several crossmodal attention blocks with layer normalization and positional embeddings. Unlike MulT, which has six crossmodal Transformers implemented in the same step, we use two crossmodal Transformers in the first step to obtain enhanced audio and vision modalities:

$$\hat{A} = CMT(V \to A) \tag{1}$$

$$\hat{V} = CMT(A \to V) \tag{2}$$

Then in the second step, another two crossmodal Transformers are used to yield the enhanced text modality representations:

$$\hat{T}_{\hat{A}} = CMT(\hat{A} \to T) \tag{3}$$

$$\hat{T}_{\hat{V}} = CMT(\hat{V} \to T) \tag{4}$$

Weighted Concatenation. After obtaining the enhanced $\hat{T}_{\hat{A}}$ and $\hat{T}_{\hat{V}}$, we concatenate them and use the self-attention to find its salient parts as the final text representation:

$$\hat{T}_{\hat{A}\hat{V}} = SA(Concat\left[\hat{T}_{\hat{A}};\ \hat{T}_{\hat{V}}\right]) \tag{5}$$

By now, crossmodal representations of every modality have been generated: $\hat{A}$, $\hat{V}$, and $\hat{T}_{\hat{A}\hat{V}}$. We concatenate them for the final representation:

$$Z = Concat\left[W_{\hat{A}}\hat{A};\ W_{\hat{V}}\hat{V};\ \hat{T}_{\hat{A}\hat{V}}\right] \tag{6}$$

where $W_{\hat{A}}$ and $W_{\hat{V}}$ are the weight matrices for audio and vision modalities, which are learned by the model itself to control how much information to extract from the two auxiliary modalities.

4.2 Automatic Primary Modality Selection – HCT-MG Model

Although we listed the literature supporting the construction of the HCT model, we hope such selection can be automatically performed by the model itself for two reasons: 1) to verify our empirical selection process is sound; 2) to ensure our proposed approach is not limited to these three modalities, but can be applied to flexible real-world scenarios where different modalities can exist, such as physiological signals. To this end, we propose the HCT-MG, an extension to improve HCT with a Modality Gating (MG) component for a modality-agnostic design. MG determines which modality should be the primary one by its learnable weights for each modality during training, rather than by manual selection. The architecture of HCT-MG is shown in Figure 5. The major difference from HCT is that MG is added between the feature projection and crossmodal Transformer components, and which modality is the primary is unknown.

5 Experiments

We describe the datasets and report our results via a comparison with prior models. Note that, during the experiments, HCT-MG automatically selected text as the primary modality, confirming the soundness of HCT and our empirical knowledge. Because MG will become obsolete once the primary modality selection converges, we frozen it at that point. We present the process by which HCT-MG selects the primary modality in the ablation study.

5.1 Datasets and Evaluation Metrics

CMU-MOSI and **CMU-MOSEI** are sentiment analysis datasets containing video clips from YouTube, annotated with sentiment scores in the range of [-3, 3]. The former has 2,199 samples, while the latter has 23,454. **IEMOCAP** is a multimodal dataset for emotion recognition. Following prior work, we use four emotions (happy, sad, angry, and neutral) for the experimental evaluation, bringing in 4,453 samples.

Following prior work on MOSI and MOSEI, we evaluate the performances using the following metrics: 7-class accuracy (i.e., Acc-7: sentiment score classification in the same scale as the labeled scores); binary accuracy (i.e., Acc-2: posi-

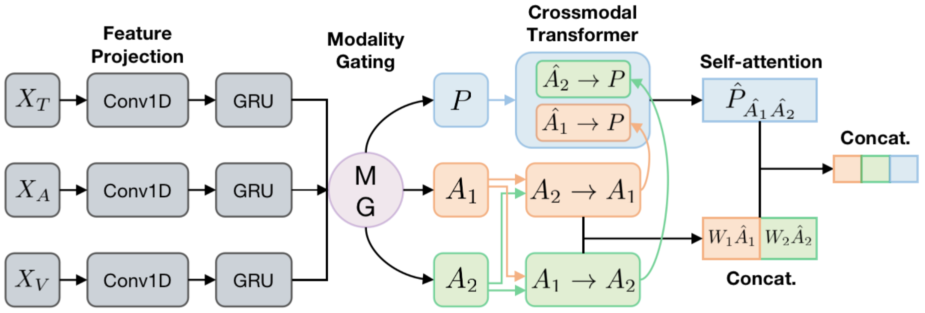


Figure 5: Architecture of HCT-MG.

tive/negative sentiment polarity); F1 score; Mean Absolute Error (MAE); and the correlation of the recognition results with ground truth. On IEMOCAP, we report the binary classification accuracy (one versus the others) and F1 score.

5.2 Experimental Evaluation

We use the features provided by the CMU-SDK (Zadeh et al., 2018c), which splits the datasets into folds for training, validation, and testing. The experimental settings are presented in the Appendix.

5.2.1 Baselines

We perform a comparative study against our approach, considering two aspects: 1) mainstream models that have been widely compared; 2) lightweight models with similar sizes to ours. Note that not every baseline has been evaluated on all three datasets. The baselines are as below:

Early Fusion LSTM (**EF-LSTM**) and Late Fusion LSTM (**LF-LSTM**) (Tsai et al., 2018). Attention or Transformer-based fusion: **RAVEN** (Wang et al., 2019), **MulT** (Tsai et al., 2019). Graph-based fusion: **Graph-MFN** (Zadeh et al., 2018b). Low-rank-based fusion: **LMF** (Liu et al., 2018). Cyclic translations-based fusion: **MCTN** (Pham et al., 2019). Context-aware attention-based fusion: **CIA** (Chauhan et al., 2019). Multi-attention Recurrent-based fusion: **MARN** (Zadeh et al., 2018c). Temporal memory-based fusion: **MFN** (Zadeh et al., 2018a). Recurrent multiple stages-based fusion: **RMFN** (Liang et al., 2018). Low-rank Transformer-based fusion: **LMF-MulT** (Sahay et al., 2020). We also include several of the above-mentioned models enhanced by Connectionist Temporal Classification (CTC). As the recognition of fine-grained emotions is significantly improved by word alignment, we do not include EF-LSTM, RAVEN, MCTN (which are tested on aligned corpora) for comparisons on IEMOCAP. Because our crossmodal Transformer component is based on MulT, we reproduced it in the same experimental environment for a fair comparison.

5.2.2 Results and Discussion

Table 2: Comparison results on CMU-MOSI and MOSEI for sentiment analysis. ↑: higher is better; ↓: lower is better; *: reproduced from the official code.

Models	CMU-MOSI				
	Acc-7↑	Acc-2↑	F1-score↑	Corr↑	MAE↓
EF-LSTM	33.7	75.3	75.2	0.608	1.023
RAVEN	33.2	78.0	76.6	0.691	0.915
MCTN	35.6	79.3	79.1	0.676	0.909
CTC+EF-LSTM	31.0	73.6	74.5	0.542	1.078
CTC+RAVEN	31.7	72.7	73.1	0.544	1.076
CTC+MCTN	32.7	75.9	76.4	0.613	0.991
MARN	34.7	77.1	77.0	0.625	0.968
MFN	34.1	77.4	77.3	0.632	0.965
RMFN	38.3	78.4	78.0	0.681	0.922
LMF	32.8	76.4	75.7	0.668	0.912
CIA	38.9	79.8	79.5	0.689	0.914
LF-LSTM(1.24M)	33.7	77.6	77.8	0.624	0.988
MulT*(1.07M)	34.3	80.3	80.4	0.645	1.008
LMF-MulT(0.84M)	34.0	78.5	78.5	0.681	0.957
HCT-MG(0.54M)	**39.4**	**82.5**	**82.5**	**0.710**	**0.881**

Models	CMU-MOSEI				
	Acc-7↑	Acc-2↑	F1-score↑	Corr↑	MAE↓
EF-LSTM	47.4	78.2	77.9	0.642	0.616
RAVEN	50.0	79.1	79.5	0.662	0.614
MCTN	49.6	79.8	80.6	0.670	0.609
CTC+EF-LSTM	46.3	76.1	75.9	0.585	0.680
CTC+RAVEN	45.5	75.4	75.7	0.599	0.664
CTC+MCTN	48.2	79.3	79.7	0.645	0.631
LMF	48.0	**82.0**	**82.1**	0.677	0.623
Graph-MFN	45.0	76.9	77.0	0.540	0.710
CIA	50.1	80.4	78.2	0.590	0.680
LF-LSTM(1.24M)	48.8	77.5	78.2	0.656	0.624
MulT*(1.07M)	50.4	80.7	80.6	0.677	0.617
LMF-MulT(0.84M)	49.3	80.8	81.3	0.668	0.620
HCT-MG(0.78M)	**50.6**	81.6	81.9	**0.691**	**0.593**

Table 3: Comparison results on IEMOCAP for emotion recognition. *: reproduced from the official code.

Models	IEMOCAP							
	Happy		Sad		Angry		Neutral	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CTC+EF-LSTM	76.2	75.7	70.2	70.5	72.7	67.1	58.1	57.4
CTC+RAVEN	77.0	76.8	67.6	65.6	65.0	64.1	**62.0**	**59.5**
CTC+MCTN	80.5	77.5	72.0	**71.7**	64.9	65.6	49.4	49.3
LF-LSTM(1.24M)	72.5	71.8	72.9	70.4	68.6	**67.9**	59.6	56.2
MulT*(1.07M)	**85.6**	**79.0**	**79.4**	70.3	**75.8**	65.4	59.5	44.7
LMF-MulT(0.86M)	**85.6**	**79.0**	**79.4**	70.3	**75.8**	65.4	59.2	44.0
HCT-MG(0.55M)	**85.6**	**79.0**	**79.4**	70.3	**75.8**	65.4	61.0	50.5

Table 4: Examples containing incongruity or ambiguity from CMU-MOSI.

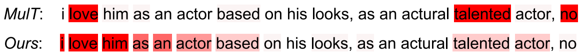
#	Spoken words + acoustic and visual behaviors	Ground truth	MulT	Ours
1	"And that's why I was not excited about the fourth one." + Uninterested tone and facial expression	-1.4	1.185	-1.416
2	"I give Shrek Forever After directed by Mike Mitchell a grade of B minus." + Smile face	1.0	-0.576	0.959
3	"Um in general um, the little kids seemed to like it that were in there." + Skeptical tone and facial expression	0.8	-1.151	0.700
4	"I honestly want the aliens to win." + Negative tone (somewhat disdainful)	-1.6	0.995	-1.906

The comparison results are shown in Table 2 and 3. On MOSI and MOSEI, it can be seen that HCT-MG improves on most metrics compared to the baselines, no matter the aligned or unaligned settings. Besides, HCT-MG outperforms the models that have similar sizes, particularly on MOSI. On IEMOCAP, we achieve the overall best performance with the same or similar results on most metrics on *Happy*, *Sad*, and *Angry*. On *Neutral*, CTC+RAVEN outperforms the others on F1 but is much worse on the other three emotions, which means it is better in distinguishing emotion from non-emotion than recognizing fine-grained emotions. Moreover, compared to LMF-MulT and MulT, which used the same crossmodal Transformer as ours, HCT-MG has the same results on *Happy*, *Sad*, and *Angry*. It is likely due to the performance bottleneck using this crossmodal Transformer without larger model sizes. However, HCT-MG outperforms them on *Neutral*, which is plausible as our hierarchical architecture works well on distinguishing all emotions.

Besides, HCT-MG was used as the crossmodal attention $(A + V) \rightarrow T$ in Exp. 2 (in Sec. 3). Figure 2 clearly shows that our approach strengthens the most salient affective parts effectively. Furthermore, we present some examples where incongruity or ambiguity exists. In Table 4 (videos available[2]), MulT fails to handle these difficult cases, producing results that contradict ground truths. In contrast, our approach can recognize true sentiments with very close numbers. The examples demonstrate that HCT-MG can successfully integrate auxiliary modalities with the primary one. We present the heatmap of these examples in the Appendix.

5.3 Ablation Study

To verify that our approach alleviates the intermodal incongruity issue and that our empirical knowledge (empirical modality selection) matches the system's (automatic modality selection), we conducted the following studies.



MulT: i love him as an actor based on his looks, as an actural talented actor, no

Ours: i love him as an actor based on his looks, as an actural talented actor, no

Figure 6: A comparison example of incongruity.

5.3.1 On Inter-Modality Incongruity

As shown in Figure 3, the affective words are enhanced by the auxiliary modalities. However, the vision modality focuses more on positive words while the audio focuses on negative words. We reimplemented MulT and extracted the attention of its enhanced text modality (attended by audio and vision) to compare with ours. In Figure 6, it can be seen that the positive and negative words are treated equally by MulT, which leaves the intermodal incongruity unsolved. This is because MulT fuses audio and vision with text at the same level and simply concatenates two enhanced text modalities. On the other hand, our approach barely gives attention to the negative word "no", showing that the incongruity is resolved at the latent level. The examples in Table 4 also demonstrate that the incongruity and ambiguity are largely resolved by our approach.

5.3.2 On Empirical Modality Selection

Although MG automatically selects text as the primary modality, which verifies our empirical knowledge, we manually change the primary modality for performance comparison using HCT. For brevity, the experiments were only conducted on MOSI. As shown in Table 5, selecting text as the primary modality achieves the best performance on every metric, once again verifying the rationality and efficacy of our empirical modality selection model.

5.3.3 On Automatic Modality Selection

As the MG automatically selects the primary modality by adjusting the weights for each modality, we show how the weights vary during training.

[2] https://sites.google.com/view/aclsubmission

Table 5: Performance comparison by manually selecting different primary modalities. $(A+V) \rightarrow T$ means Text is the primary. ↑: higher is better; ↓: lower is better;

Fusion	Acc-7↑	Acc-2↑	F1↑	Corr↑	MAE↓
$(A+V) \rightarrow T$	**38.9**	**82.5**	**82.6**	**0.717**	**0.859**
$(T+V) \rightarrow A$	37.5	81.3	81.3	0.705	0.883
$(A+T) \rightarrow V$	38.3	80.9	81.0	0.679	0.909

The weight of a modality denotes the probability that this modality is selected as the primary one. Figure 7 shows how the probabilities of the modalities vary in the first epoch. It can be noted that the text modality is not the primary one at the beginning but gradually dominates after batch 60 during the training. Figure 8 shows the variation in the average probability of each modality with epoch during training. It can be seen that the text modality does dominate, and the probability distribution starts to converge at around 40 epochs. The results of Figure 7 and Figure 8 are consistent with Table 5. It confirms again that our empirical knowledge is rational on the three benchmarks based on text, audio, and vision modalities. Furthermore, now that we know how HCT-MG works, its utility is expected to expand beyond trimodal settings and into general scenarios with different and more signals (e.g., Electroencephalogram (EEG), Electrocardiogram (ECG), Galvanic Skin Response (GSR)).

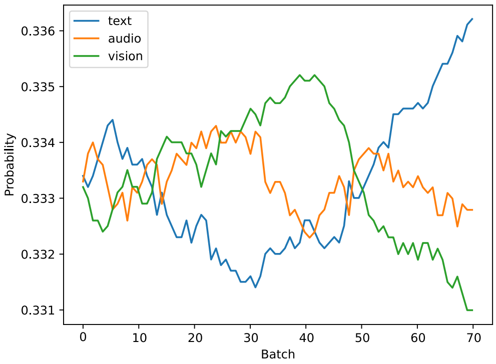


Figure 7: Probability (weight) variation of each modality in the first epoch.

6 Conclusions

In this work, we conduct a thorough analysis on crossmodal attention-based multimodal fusion and propose a hierarchical crossmodal Transformer with modality gating for multimodal sentiment and emotion recognition. The major contributions are:

1) We demonstrate the existence of inter-modal incongruity at the latent level due to crossmodal attention. Specifically, via step-by-step visualization analysis, we show that crossmodal attention can help capture affective information across modalities and strengthen salient parts in the target modality, but it can also bring mismatched affective tendencies from different modalities.

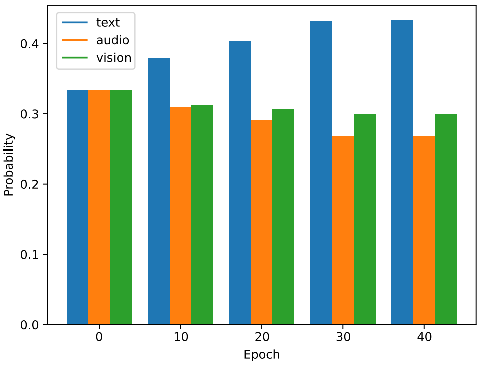


Figure 8: Average probability (weight) variation of each modality with epoch.

2) We gather previous studies as empirical knowledge that supports the selection of text as the primary modality to build a hierarchical crossmodal Transformer–HCT, which requires fewer fusions and does not fuse a single modality repeatedly. An ablation study demonstrates that manually using text as the primary modality yields better results than using audio and vision.

3) We incorporate Modality Gating (MG) into HCT for an extension version–HCT-MG, which can automatically select the primary modality during training. The selection process confirms the soundness of our empirical knowledge and supports the rationality of previous studies. HCT-MG alleviates the inter-modal incongruity and reduces the model size to less than 1M while significantly outperforming existing models of similar sizes. We believe that the use of HCT-MG is not limited to trimodal sentiment and emotion analysis, but can be extended to more affective computing tasks.

In our future work, we will test HCT-MG in other domains, especially where the affective tendencies of different modalities easily mismatch, such as humor and sarcasm detection (Pramanick et al., 2022), and where other modalities exist, such as biological signals (Li et al., 2022b). We will also try some principle component analysis techniques (Shao et al., 2022) to further reduce redundant and misleading information by removing irrelevant attributes while retaining the primary ones. We expect HCT-MG to be used for solving multimodal affective computing tasks in real life.

Limitations

Although the inter-modal incongruity is largely removed by the hierarchical architecture, the affective tendency could be wrong if ambiguity exists in the primary modality. In Figure 6, our approach recognizes this sample as positive with the score of *+1.84*, but the ground truth is labeled as negative with the score of *-1*. It is likely because the first half of the text denotes a positive sentiment, yet the second is obviously negative. Without contextual knowledge, it is almost impossible for a system to know that the second half is the focus of the content, as humans can. As our model uses Glove as text features, which are non-contextual word embeddings, it is hard to capture the semantic meaning. This problem can likely be resolved with contextual language models (e.g., BERT) or auxiliary modalities that have the same affective tendency (which can modify the primary modality).

Also, as with all the other supervised learning tasks, our model relies on the accuracy of the labels. However, it is uneasy to label difficult cases where inter-modal incongruity, ambiguous emotions, or missing information exist. Take the cases in Table 4 as an example: 1) In our opinion, #2 should be a neutral emotion with the value of 0, as the person is just stating a fact. 2) There is a word "but" missing at the end of the labeled sentence of #3 (can be clearly noticed in the audio or the video), which is a sign indicating a turnaround in attitude. It can be regarded as the same situation as the sample in Figure 6. However, #3 separates the whole sentence into two sub-sentences, while the sample in Figure 6 combines two sub-sentences as a whole. Such an inconsistency in labeling introduces incongruity and ambiguity into the tasks and hinders the training of robust and applicable models.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629.

Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5647–5657.

Ze-Jing Chuang and Chung-Hsien Wu. 2004. Multi-modal emotion recognition from speech and text. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing*, pages 45–62.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.

Ken W Grant and Steven Greenberg. 2001. Speech intelligibility derived from asynchronous processing of auditory-visual information. In *AVSP 2001-International Conference on Auditory-Visual Speech Processing*.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956.

Yuanchao Li. 2018. Towards improving speech emotion recognition for in-vehicle agents: Preliminary results of incorporating sentiment analysis by using early and late fusion methods. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 365–367.

Yuanchao Li, Peter Bell, and Catherine Lai. 2022a. Fusing ASR outputs in joint training for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7362–7366. IEEE.

Yuanchao Li, Peter Bell, and Catherine Lai. 2022b. Multimodal dyadic impression recognition via listener adaptive cross-domain fusion. *arXiv preprint arXiv:2211.05163*.

Yuanchao Li, Carlos Toshinori Ishi, Koji Inoue, Shizuka Nakamura, and Tatsuya Kawahara. 2019. Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction. *Advanced Robotics*, 33(20):1030–1041.

Yuanchao Li, Tianyu Zhao, and Xun Shen. 2020. Attention-based multimodal fusion for estimating human emotion in real-world HRI. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 340–342.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161.

Kristen A Lindquist, Jennifer K MacCormack, and Holly Shablack. 2015. The role of language in emotion: Predictions from psychological constructionism. *Frontiers in psychology*, 6:444.

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. *arXiv preprint arXiv:1806.00064*.

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.

Jonathan E Peelle, Ingrid Johnsrude, and Matthew H Davis. 2010. Hierarchical processing for speech in human auditory cortex and beyond. *Frontiers in human neuroscience*, page 51.

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.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Shraman Pramanick, Aniket Roy, and Vishal M Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3930–3940.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Vandana Rajan, Alessio Brutti, and Andrea Cavallaro. 2022. Is cross-attention preferable to self-attention for multi-modal emotion recognition? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4693–4697. IEEE.

Ahmed Rashed, Shereen Elsayed, and Lars Schmidt-Thieme. 2022. Context and attribute-aware sequential recommendation via cross-attention. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 71–80.

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

Shun Shao, Yftah Ziser, and Shay B Cohen. 2022. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. *arXiv preprint arXiv:2203.07893*.

Leimin Tian, Johanna Moore, and Catherine Lai. 2016. Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 565–572. IEEE.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 881–888.

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.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

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.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Yang Wu, Yanyan Zhao, Xin Lu, Bing Qin, Yin Wu, Jian Sheng, and Jinlong Li. 2021. Modeling incongruity between modalities for multimodal sarcasm detection. *IEEE MultiMedia*, 28(2):86–95.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 929–932.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

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 *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. 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.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018c. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2019. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Shucong Zhang, Malcolm Chadwick, Alberto Gil CP Ramos, and Sourav Bhattacharya. 2022. Cross-attention is all you need: Real-time streaming transformers for personalised speech enhancement. *arXiv preprint arXiv:2211.04346*.

A Appendix

A.1 Datasets

Table 6: Data distribution and modality sampling rate of CMU-MOSI and CMU-MOSEI. S_A for audio sampling rate and S_V for vision sampling rate.

Dataset	**Train**	**Valid**	**Test**	**Total**	S_A	S_V
CMU-MOSI	1284	229	686	2199	12.5	15
CMU-MOSEI	16,326	1871	4659	22,856	20	15

Table 7: Data distribution of four emotions in the IEMOCAP dataset.

Emotions	**Train**	**Valid**	**Test**	**Total**
Neural	954	358	383	1695
Happy	338	116	135	589
Sad	690	188	193	1071
Angry	735	136	227	1098
Total	2717	798	938	4453

Table 8: Sequence lengths and feature dimensions of the three modalities in the three benchmark datasets. *: The development team screened the vision and audio features of CMU-MOSI.

Dataset	**Text**		**Vision**		**Audio**	
	len	dim	len	dim	len	dim
CMU-MOSI	50	300	500	20*	375	5*
CMU-MOSEI	50	300	500	35	500	74
IEMOCAP	20	300	500	35	400	74

A.2 Extracted Features

The sequence lengths and feature dimensions of the three modalities in the three benchmarks are shown in Table 8.

Textual Features: GloVe. The transcriptions in all three datasets use the global word embeddings generated by GloVe. This distributed representation allows words in the same context to be close to each other in the vector space and maintain specific relationships (Pennington et al., 2014). For this pre-extracted data, the text modal features are trained and derived from 840 billion tokens with 300 dimensions of GloVe embeddings.

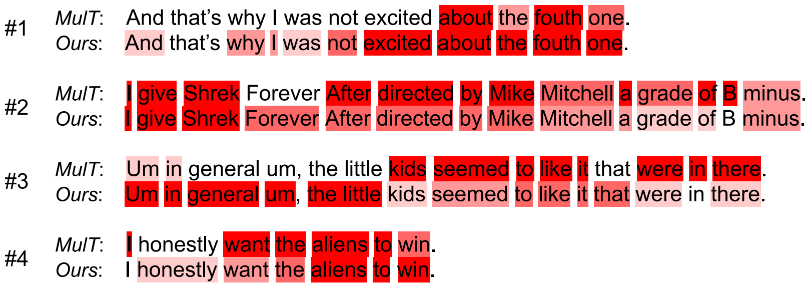
Vision Features: FACET. FACET is a commercial facial emotion detection software developed by iMotions[3]. The software can demonstrate 35 facial action units and record facial muscle movements to represent frame-by-frame emotions.

Audio Features: COVAREP. COVAREP is an open-source repository for speech processing, supporting collaboration and free access. The features of the processed speech data are based on pitch tracking, polarity detection, spectral envelopes, glottal flow, and other common speech features (Degottex et al., 2014). The pre-extracted data contains 74 dimensions of speech features.

A.3 Hyperparameters Tuning

After tuning the hyperparameters, we find the optimal settings, as shown in Table 9.

[3] https://imotions.com/platform/



#1 *MulT*: And that's why I was not excited about the fouth one.
Ours: And that's why I was not excited about the fouth one.

#2 *MulT*: I give Shrek Forever After directed by Mike Mitchell a grade of B minus.
Ours: I give Shrek Forever After directed by Mike Mitchell a grade of B minus.

#3 *MulT*: Um in general um, the little kids seemed to like it that were in there.
Ours: Um in general um, the little kids seemed to like it that were in there.

#4 *MulT*: I honestly want the aliens to win.
Ours: I honestly want the aliens to win.

Figure 9: Heatmap comparison of the examples in Table 4 (Page 7).

Table 9: Hyperparameter settings for the three datasets.

Setting	**CMU-MOSI**	**CMU-MOSEI**	**IEMOCAP**
learning rate	1e-3	1e-3	1e-5
batch size	36	64	16
hidden size	40	40	40
kernel (T/A/V)	1/1/1	1/1/1	1/1/1
decay when	20	20	20
number of epochs	30	30	60
transformer layers	2	4	2
attention heads	5	5	5

A.4 Heatmap Comparison

The heatmap values are from the enhanced text modalities (attended by audio and vision) of MulT and ours. With the the hierarchical architecture of our approach, some words that are not highlighted in MulT are highlighted.