CLASP: Cross-modal Alignment Using Pre-trained Unimodal Models

Jianing Zhou[1], Ziheng Zeng[1], Hongyu Gong[2], Suma Bhat[1]
[1]University of Illinois Urbana-Champaign
[2]Facebook AI
[1]{zjn1746,zzeng13,spbhat2}@illinois.edu
[2]hygong@fb.com

Abstract

Recent advancements in joint speech-text pre-training have significantly advanced the processing of natural language. However, a key limitation is their reliance on parallel speech-text data, posing challenges due to data accessibility. Addressing this, our paper introduces an innovative framework for jointly performing speech and text processing *without* parallel corpora during pre-training but only downstream. Utilizing pre-trained unimodal models, we extract distinct representations for speech and text, aligning them effectively in a newly defined space using a multi-level contrastive learning mechanism. A unique swap reconstruction mechanism enhances the alignment and is followed by fusion via a multi-head mechanism, seamlessly merging modality-invariant and modality-specific representations. Testing for emotion recognition (Spoken Language Understanding task) and idiom usage detection (Natural Language Understanding task) demonstrates robust performance, with commendable robustness to noise in text or speech data.

1 Introduction

In recent years, advancements in speech-text pre-training for learning universal feature representations from large training corpora have been significant (Bapna et al., 2021; Li et al., 2021; Tang et al., 2022), leading to notable success in various unimodal (Lin and Xu, 2019; Zhang et al., 2022) and multimodal downstream tasks (Busso et al., 2008; Zadeh et al., 2016, 2018). These approaches leverage multimodal self-supervised learning objectives, including cross-modal masked data modeling (Li et al., 2021; Kang et al., 2022) and cross-modal contrastive learning (Sachidananda et al., 2022), aimed at aligning speech representations with their corresponding text representations. However, a primary drawback of these methods is their reliance on *parallel* speech-text data *during* the pre-training phase.

Unlike non-parallel unimodal corpora collections, acquiring parallel corpora (e.g., datasets like LibriSpeech (Panayotov et al., 2015)) often requires extensive manual filtering and annotation efforts. This necessity significantly limits the amount of parallel corpora compared to their unimodal counterparts, particularly in modality-scarce scenarios, such as when data is available in only one modality (e.g., languages that are primarily spoken). Our paper addresses this limitation.

Our approach involves leveraging unimodal models pre-trained on non-parallel corpora to obtain speech and text representations, subsequently aligning them using a small sample of task-specific parallel data. This strategy not only reduces the dependence on large parallel corpora during pre-training but also enables the utilization of extensive and diverse repositories of unimodal data, expanding the effectiveness and applicability of speech and text pre-training methods.

Our study introduces a framework (CLASP) for processing multimodal information in speech and text by learning both factorized and shared representations for each modality, followed by fusing the learned representations with only parallel data from downstream tasks. Inspired by previous attempts to disentangle representations from different modalities into modality-specific and modality-invariant components (Hazarika et al., 2020), and innovating beyond solely using contrastive learning at the modality level; we incorporate a multi-level contrastive learning mechanism for representation disentanglement. Additionally, we propose a novel swap reconstruction mechanism based on the intuition that ideal modality-invariant and modality-specific representations should be capable of reconstructing their original unimodal representations when combined. This mechanism promotes tighter cross-modal alignment by aligning modality-invariant representations between text and speech within each batch of data. Acknowledg-

Findings of the Association for Computational Linguistics: ACL 2024, pages 11518–11531

ing the significance of modality-specific representations for certain tasks, such as emotion recognition (e.g., prosodic elements for emotion (Hazarika et al., 2020)), and the varied roles of different heads in pretrained models of different modalities, we integrate a multi-head fusion mechanism to merge modality-specific and modality-invariant representations.

CLASP's efficacy is validated through extensive experiments on different language understanding tasks that benefit from combining the speech and text modalities. We experiment with two spoken language understanding (SLU) tasks with parallel text input (emotion recognition (Busso et al., 2008) and intent classification (Bastianelli et al., 2020)) and an natural language understanding (NLU) task with parallel speech input. The latter, which we call multimodal idiom usage detection, is novel (innovating over text-only idiomatic/literal disambiguation previosuly studied (Zeng and Bhat, 2021; Zhou et al., 2023) and aims to differentiate between literal and figurative uses of idioms using *both* text and speech data. This task operates at a more granular level than emotion detection and is significant because many emotions are expressed through abstract feelings conveyed using figurative expressions (Gibbs Jr et al., 2002; Glucksberg and McGlone, 2001; Fussell and Moss, 2014). Our approach results in consistent gains over several competent multimodal methods for both tasks.

Overall, the main contributions are as follows:

- We propose CLASP, a framework for learning multimodal representations from pre-trained unimodal models for tasks involving speech and text without requiring pre-training on parallel speech-text corpora.
- We conduct the first study on *multimodal idiom usage detection* using both speech and text, by creating a dataset consisting of 6,325 annotated text-speech pairs. Our dataset will be available at `https://github.com/zhjjn/CLASP.git`.
- Evaluating our proposed framework across three tasks in multimodal settings—the classical emotion recognition, intent classification (text enhancing the SLU task) and our newly proposed idiom usage detection task (a less explored area where speech enhances the NLU task)—confirm the effectiveness and generality of our framework.
- CLASP demonstrates exceptional resilience to noisy multimodal data, achieving significant improvements over prior models. This robustness is crucial in modality-scarce scenarios (e.g., languages with mostly speech data), where the missing modality may be automatically generated (potentially noisily) from available unimodal data. Detailed ablation studies and analyses further substantiate our claims.

2 Related Work

Multimodal Pre-training with Speech and Text. Compared to the extensive explorations in the realm of multimodal pre-training for vision-and-text (Radford et al., 2021; Li et al., 2022a,b), the area of speech-text pre-training remains underexplored. A few notable works in this domain include SpeechBERT (Chuang et al., 2020), CTAL (Li et al., 2021), SLAM (Bapna et al., 2021), ST-BERT (Kim et al., 2021), CALM (Sachidananda et al., 2022), Maestro (Chen et al., 2022) and STPT (Tang et al., 2022), which focus on the joint training of speech and text with aligned data in the pretraining corpus, including those in low-resource data environments (Kang et al., 2022) and conversational settings that improve modeling the contextual information (Yu et al., 2023). All these works rely on aligning the modalities during large-scale pre-training utilizing parallel data from speech and text. In contrast, our work utilizes unimodal pre-trained models for speech and text, but relies on aligning the modalities using task-specific parallel data (potentially in smaller numbers compared to the large pre-training data), thereby achieving the multimodal alignment in a task-specific finetuning setting.

Multimodal Fusion. In addition to the body of research on multimodal pre-training summarized above, the realm of modality fusion strategies has seen advances, particularly focusing on the integration of features extracted from unimodal modules. A few notable examples are works focusing on directly fusing extracted multimodal representations including the Tensor Fusion Network (TFN) presented by Zadeh et al. (2017) and Low-rank Multimodal Fusion (LMF) (Liu et al., 2018), works focusing on first disentangling multimodal representations and then fusing them together (Hazarika et al., 2020; Yang et al., 2022), works utilizing mutual information maximization (Han et al., 2021;

Colombo et al., 2021) and works focusing on fusion of multimodal representations from uni-modal transformers including LFMIM (Sun et al., 2023). However, previous research in this area performs contrastive learning at the modality level, ignoring the semantic level constrasts and overlooking the importance of multi-level disentanglement, encompassing both semantic and modality levels. In contrast, our research introduces a novel swap reconstruction mechanism that innovatively combines these two crucial aspects. Additionally, little is known about the extent to which the synthesis of semantic-related and modality-related information can replace the missing modality, which we centrally study in our experiments.

3 Framework

Our framework combines unimodal representations from pre-trained speech and text models into a multimodal representation in a fine-tuning stage for downstream tasks. This eliminates the necessity for pre-training on parallel speech-text data, and that of extensive parallel data. The functioning of our proposed framework can be divided into three main stages: (1) Disentangling multimodal representation using a multi-level contrastive learning mechanism as previous works only perform contrastive learning on the modality level but ignoring the semantic level whereas both levels are important; (2) cross-modal alignment through a swap reconstruction mechanism which reconstructs the unimodal encoding from the shared modality-invariant encoding and modality-specific encoding from opposite modalities because the ideal modality-invariant and modality-specific representations should be capable of reconstructing their original unimodal representations when combined; and (3) Fusing multimodal representation via a multi-head fusion mechanism to pay different attentions to different heads for different modalities. Figure 1 illustrates the workflow of our proposed framework and its details follow.

Unimodal Representations. We utilize unimodal pre-trained models to extract representations—RoBERTa-base (Liu et al., 2019) for text and wav2vec2-base (Schneider et al., 2019) for speech:

$$\mathbf{r}_a^i = \text{wav2vec2}(\mathbf{X}_a^i)$$
$$\mathbf{r}_t^i = \text{RoBERTa}(\mathbf{X}_t^i),$$

where $\mathbf{X}_a^i$ and $\mathbf{X}_t^i$ are the speech data and text data respectively, and $\mathbf{r}_a^i$ and $\mathbf{r}_t^i$ refer to the speech representations and text representations correspondingly. Next, we disentangle, align, and fuse these two representations as detailed below.

Here we briefly define notations used in the following model formulation. For the representation, the subscripts denote the modality (a for speech and t for text) and whether the representation is invariant (inv) or specific (spe) to the modality. Thus, $\mathbf{h}_{a,\text{inv}}^i$, stands for a representation in the speech modality in the modality-invariant setting.

3.1 Multi-Level Contrastive Learning for Disentanglement

Once we have the unimodal speech and text representations, we disentangle each into two separate representations: the *modality-invariant* and the *modality-specific* representations. The modality-invariant representation captures the shared semantics between modalities. Ideally, the modality-invariant components extracted from parallel unimodal representations should be identical. Thus, for a given example, we first use a linear layer to learn the hidden modality-invariant representations from the speech and text unimodal representations:

$$\mathbf{h}_{a,\text{inv}}^i = E_{\text{inv}}(\mathbf{r}_a^i), \quad \mathbf{h}_{t,\text{inv}}^i = E_{\text{inv}}(\mathbf{r}_t^i).$$

To ensure consistency across modalities, we treat modality-invariant representations derived from matching speech and text examples as positive pairs. Conversely, modality-invariant representations from different examples within the same batch are considered negative pairs. We then apply contrastive learning at the semantic level to learn these modality-invariant representations.

$$\mathcal{L}_{\text{cts}}^{\text{sem}} = -\log \frac{f(\mathbf{h}_{a,\text{inv}}^i, \mathbf{h}_{t,\text{inv}}^i)}{\sum_{m\in\{a,t\}} \sum_{j\neq i} f(\mathbf{h}_{m,\text{inv}}^i, \mathbf{h}_{m,\text{inv}}^j)} \tag{1}$$

In contrast, modality-specific representations only contain information unique to each modality, without encoding any semantics. Therefore, given the text and speech representations for the same example, we employ two distinct linear layers to extract their modality-specific representations:

$$\mathbf{h}_{a,\text{spe}}^i = E_{\text{a}}(\mathbf{r}_a^i), \quad \mathbf{h}_{t,\text{spe}}^i = E_{\text{t}}(\mathbf{r}_t^i)$$

To ensure that the modality-specific representation for the same modality contains consistent information across different examples, we treat the extracted modality-specific representation for the

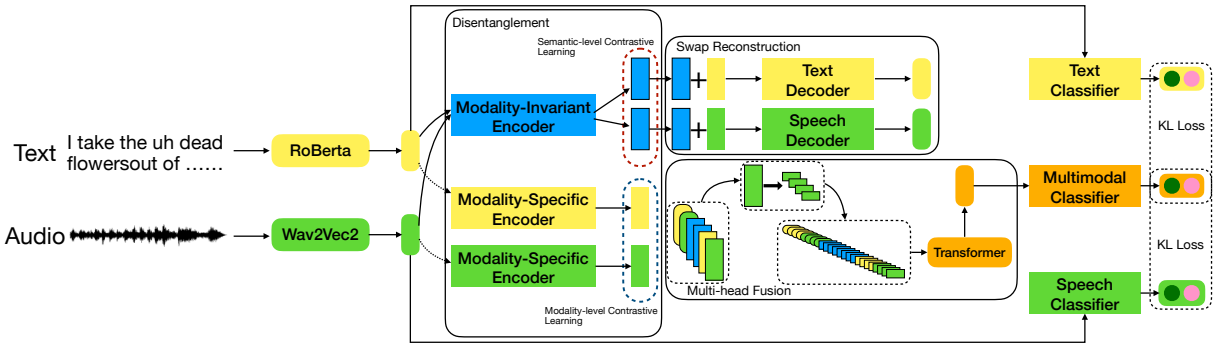


Figure 1: The overview of our framework.

same modality as positive pairs and cross-modality ones as negative pairs. We then apply contrastive learning at the modality level to learn the modality-specific speech and text representations.

$$\mathcal{L}_{\text{cts}}^{\text{mod}} = \log \frac{f(\mathbf{h}_{a,\text{spe}}^{i}, \mathbf{h}_{t,\text{spe}}^{i})}{\sum_{m\in\{a,t\}} \sum_{j\neq i} f(\mathbf{h}_{m,\text{spe}}^{i}, \mathbf{h}_{m,\text{spe}}^{j})} \tag{2}$$

In the end, each unimodal speech and text representation is disentangled into two parts: a modality-invariant representation that solely encodes semantics, and a modality-specific representation that captures the modality signatures. This results in four distinct representations.

3.2 Swap Reconstruction Mechanism

We enhance the quality of both modality-invariant and modality-specific representations derived from each modality by employing a reconstruction mechanism. Our assumption is that the ideal modality-invariant and modality-specific representations should be capable of reconstructing their original unimodal representations when combined, i.e., the fusion of modality-invariant representation from a single modality with its modality-specific representation should yield that modality's original representation.

This mechanism assumes that modality-invariant representations from various modalities encapsulate the same semantic information, while modality-specific representations hold unique modality-related information. Together, we devise a swap reconstruction process that uses modality-invariant and modality-specific representations from different modalities to reconstruct their original representations. For instance, we reconstruct the original unimodal speech representations by merging the text modality-invariant representation with the speech modality-specific representation:

$$\hat{\mathbf{r}}_{a}^{i} = \hat{E}_{a}(\mathbf{h}_{a,\text{spe}}^{i} + \mathbf{h}_{t,\text{inv}}^{i})$$

where $\hat{E}_a$ is a linear layer. A parallel process is applied for text representations, wherein semantic information from speech-derived modality-invariant representations is combined with modality-specific representation from text:

$$\hat{\mathbf{r}}_{t}^{i} = \hat{E}_{t}(\mathbf{h}_{t,\text{spe}}^{i} + \mathbf{h}_{a,\text{inv}}^{i})$$

where $\hat{E}_t$ is a linear layer. Finally, we compute the reconstruction loss as the mean squared error between the original and the reconstructed representations:

$$\mathcal{L}_{recon} = \frac{1}{2}\Big(\sum_{m\in\{a,t\}} \frac{\| \mathbf{r}_{m}^{i} - \hat{\mathbf{r}}_{m}^{i} \|}{d_h}\Big) \tag{3}$$

where d_h is the hidden size.

3.3 Multi-Head Fusion Mechanism

Next, we propose a multi-head fusion mechanism to synthesize a cohesive joint speech-text representation which is pivotal for downstream tasks. Traditional fusion methodologies in this domain typically employ self-attention mechanisms, rooted in the Transformer architecture, to merge a concatenation of the original, modality-invariant, and modality-specific representations (Hazarika et al., 2020). Towards this end, leveraging the self-attention mechanism based on the Transformer model, our approach innovatively incorporates a multi-head fusion mechanism.

The multi-head attention operationalizes each head with distinct key, query, and value transformation matrices. In this setup, the token representation transforms into key, query, and value vectors through these matrices. Subsequently, each

attention head calculates the degree of attention this token should allocate over the entire input sequence, which is based on the correlation between its query vector and the key vectors of other tokens. The value vectors, modulated by these attention weights, constitute the contextualized representation of the token. These are then subject to a linear projection, forming the output of each attention head. The outputs from all heads are then concatenated to yield the final integrated output:

$$\mathbf{r}_a^i = \mathbf{r}_a^{i,(1)} \oplus \cdots \oplus \mathbf{r}_a^{i,(h)} \oplus \cdots \oplus \mathbf{r}_a^{i,(H)}$$

where H is the number of attention heads and $\oplus$ denotes the vector concatenation. Given that our projected and original representations are derived directly from pre-trained unimodal models, whose outputs are concatenated from all heads, they enter the fusion stage as fully formed entities while losing focus on different heads.

A critical insight driving our approach is that uniformly assigning attention scores across different heads within the same representation may not be ideal. This is because the heads may capture distinct patterns and features, which could have a differential impact on downstream tasks. To account for this, our multi-head fusion mechanism initially dissects each representation into separate outputs based on their respective heads:

$$\mathbf{r}_m^i \Rightarrow \{\mathbf{r}_m^{i,(1)}, \cdots, \mathbf{r}_m^{i,(h)}, \cdots, \mathbf{r}_m^{i,(H)}\}, m \in \{a, t\}$$
$$\mathbf{h}_{m,\mathrm{s}}^i \Rightarrow \{\mathbf{h}_{m,\mathrm{s}}^{i,(1)}, \cdots, \mathbf{h}_{m,\mathrm{s}}^{i,(h)}, \cdots, \mathbf{h}_{m,\mathrm{s}}^{i,(H)}\}, m \in \{a, t\}, \mathrm{s} \in \{\mathrm{inv}, \mathrm{spe}\}$$

These are then assembled into a matrix:

$$\mathbf{M}^i = [\mathbf{r}_a^i, \mathbf{r}_t^i, \mathbf{h}_{a,\mathrm{inv}}^i, \mathbf{h}_{t,\mathrm{inv}}^i, \mathbf{h}_{t,\mathrm{spe}}^i, \mathbf{h}_{t,\mathrm{spe}}^i]$$

over which a multi-head self-attention is performed. This procedure ensures that each representation becomes aware of the other cross-modal, cross-subspace, and cross-head representations. The outputs from this fusion process are then concatenated to form the input for the final multimodal classifier.

3.4 Learning

The overall learning of the model is governed by the following losses.

Task Loss. The task-specific loss, used to estimate prediction quality during training, employs the standard cross-entropy loss. This loss is calculated using output logits from three linear-layer classifiers: a text classifier, a speech classifier, and a multimodal classifier, with their respective losses $\mathcal{L}_\mathrm{t}$, $\mathcal{L}_\mathrm{a}$, and $\mathcal{L}_\mathrm{m}$. The text classifier uses the original text representations from RoBERTa; the speech classifier the original speech representations from wav2vec2, and the multimodal classifier our fused multimodal representations. Therefore, the task loss is calculated as follows:

$$\mathcal{L}_\mathrm{task} = \mathcal{L}_\mathrm{t} + \mathcal{L}_\mathrm{a} + \mathcal{L}_\mathrm{m}$$

KL Loss. Ideally, the three sets of logits from the above classifiers, namely text, speech, and multimodal, should have similar distributions since they correspond to the same ground truth labels. To estimate and minimize the distribution distance between the text and multimodal logits, as well as the speech and multimodal logits, we employ two KL-divergence losses:

$$\mathcal{L}_\mathrm{kl} = \mathcal{L}_\mathrm{kl}^\mathrm{t,m} + \mathcal{L}_\mathrm{kl}^\mathrm{a,m}$$

Contrastive Loss. $\mathcal{L}_\mathrm{cts}^\mathrm{sem}$ and $\mathcal{L}_\mathrm{cts}^\mathrm{mod}$ represent the loss corresponding to our multi-level constrastive learning, which are described in equations 1 and 2.

Reconstruction Loss. $\mathcal{L}_\mathrm{recon}$ represents the loss corresponding to our swap reconstruction mechanism, described in equation 3.

Finally, the overall loss function is the sum of the above losses, which we minimize during model training.

$$\mathcal{L} = \mathcal{L}_\mathrm{task} + \alpha\mathcal{L}_\mathrm{cts}^\mathrm{sem} + \beta\mathcal{L}_\mathrm{cts}^\mathrm{mod} + \gamma\mathcal{L}_\mathrm{recon} + \delta\mathcal{L}_\mathrm{kl}$$

4 Idiom Usage Detection

When individuals articulate their emotions verbally, they delve into a somewhat abstract realm, often prompting the use of figurative language (Gibbs Jr et al., 2002; Fussell and Moss, 2014). Thus, a precursor to detecting emotions is to detect the occurrence of figurative language, typically done using idiomatic expressions. Here, we study this in the specific context of the use of idioms, which can be in their literal sense or idiomatic sense in a context-dependent manner (Sag et al., 2002). Our key insight is that detecting whether an idiom is used in a literal or figurative sense from text (see Table 1) is aided by accessing the corresponding utterance and its prosodic features (e.g., stress patterns), among other speech attributes. For instance, in the sentence, “Relax, I’m just pulling your leg!" the prosody of the latter part can help label the idiom as figurative.

Input	He can **blow this trumpet** after success .
Output	**Figurative**
Input	A band **blew** long brass **trumpets** .
Output	**Literal**

Table 1: Examples of input and output for the idiom usage detection task. Idioms used figuratively are highlighted in **bold red**. Idioms used literally are highlighted in **bold blue**.

Statistics	Idiomatic	Literal	All
Number of Idioms	711	613	1107
Avg. length of idioms	2.67	2.42	2.60
Number of instances	1,733	4,592	6,325
Avg. length of instance	36.77	37.65	37.41
Avg. duration of instance	13.25	13.33	13.31

Table 2: Statistics of our parallel corpus.

Here, we detail the process of data collection and annotation of the Multimodal Idiom uSage deTection dataset (MIST) for this task.

4.1 Data Collection

We developed a comprehensive multimodal dataset for idiom usage detection based on the widely recognized LibriSpeech dataset (Panayotov et al., 2015). We curated text-speech pairs, each containing an idiomatic expression, to support our goal of detecting literal and figurative idiom usage from textual and spoken samples.

The methodology for assembling text-speech pairs is inspired by the approach used in constructing the MAGPIE dataset, as detailed in Haagsma et al. (2020). We begin by selecting a comprehensive set of idioms from Wiktionary, given its expansive scope and diverse coverage of idiomatic expressions. Subsequently, we employed the pre-extraction system delineated by Haagsma et al. (2019) to systematically extract all variants of these idioms from our base corpus, the LibriSpeech dataset transcriptions. This method identified approximately 13,000 instances from the 100-hour LibriSpeech dataset. However, to reduce the complexity in the idiom usage detection task, we filter out instances containing multiple idioms, retaining only those with a *single* idiom.

The data collection process ultimately yielded a corpus comprising 1,107 distinct idiomatic expressions, appearing in a total of 6,325 text-speech pairs (see Table 2 for detailed statistics).

4.2 Data Annotation

A crucial aspect in the creation of the multimodal idiom dataset is the accurate annotation of idioms for their literal or figurative usage. To address this need, we employed a state-of-the-art neural model for idiom usage detection with an accuracy over 0.96, referred to as CLCL (Zhou et al., 2023), to automatically generate preliminary labels for each text instance. The model processes the provided transcriptions and the list of idioms, subsequently classifying each as either literally or figuratively used. As demonstrated by Zhou et al. (2023), the capabilities of the CLCL model extend to recognizing the usage of idioms not encountered in training data. This feature is particularly beneficial for applying the model to detect the usage of the diverse idioms found in the LibriSpeech dataset.

4.3 Corpus Analysis

We summarize the statistics of the newly constructed MIST dataset in Table 2. Total duration of the entire dataset is 23.28 hours. There are in total 6,325 audio segments of literal and figurative instances. The average duration of each data instance is 13.31 seconds, with idiomatic instance an average of 13.25 seconds and literal instance with an average of 13.33 seconds. The average number of words per idiom is 2.6 and the average number of words in sentences is 37.41, sentences with idiomatic instance had an average of 36.77 words and those with a literal instance had an average of 37.65 words. Therefore, both audio duration and sentence length cannot be used to infer the usage of idiomatic or literal.

5 Experiments

5.1 Dataset

We conduct experiments on the popular multimodal task of emotion recognition using the IEMOCAP dataset (Busso et al., 2008), intent recognition using the SLURP dataset (Bastianelli et al., 2020) and our proposed idiom usage detection using the purposefully created MIST dataset (see Section 4). For the IEMOCAP dataset, we adhered to the settings outlined in (Kang et al., 2022) and perform a 4-way classification (referred as IEMOCAP4). Note that our experiments were conducted solely at the utterance level for emotion recognition, rather than at the conversation level.

Task	Given Example	Truth	Text	Speech	Ours
IEMOCAP4	You're the only one I know who loves his parents.	Neutral	Sad	Sad	Neutral
IEMOCAP4	You infuriate me sometimes. Do you know that? God.	Angry	Neutral	Happy	Angry
MIST	we had just left it outside and were all **on fire** to get back to it	Figurative	Literal	Literal	Figurative
MIST	then she heaved a sigh and **wiped her eye** and ran o'er hill	Literal	Figurative	Figurative	Literal

Table 3: Qualitative Examples

Methods	Roberta	Wav2vec2	LFMIM	Ours
Acc	$0.86^{+.005}_{-.005}$	$0.72^{+.005}_{-.002}$	$0.87^{+.01}_{-.005}$	$\mathbf{0.88}^{+.005}_{-.003}$
F1	$0.85^{+.01}_{-.01}$	$0.71^{+.01}_{-.005}$	$0.86^{+.003}_{-.005}$	$\mathbf{0.87}^{+.005}_{-.002}$

Table 4: Performance on SLURP. Numbers in superscript and footscript represent the 95% confidence interval based on significance test.

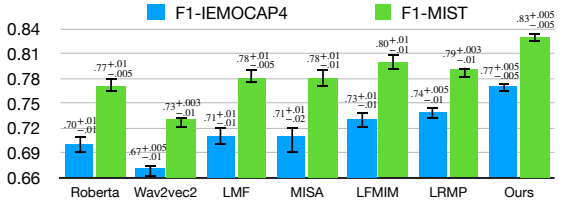


Figure 2: Performance of different methods on two benchmark datasets. Difference between different models is significant based on our statistic significance test. Numerical results are presented in the appendix.

5.2 Baselines

We select state-of-the-art multimodal emotion recognition models—LMF (Liu et al., 2018), MISA (Hazarika et al., 2020) and LFMIM (Sun et al., 2023)— as baseline models.

Additionally, we compare three types of unimodal pre-trained models, including Wav2vec2-base model (Schneider et al., 2019) for speech data, RoBERTa-base model (Liu et al., 2019) for only text data, and a single speech-text multimodal pre-trained model, LRMP (Kang et al., 2022).

5.3 Experimental Settings

We utilize wav2vec2-base model as our speech encoder and a RoBERTa-base model as our text encoder. For our multi-head fusion mechanism, we utilize a Transformer encoder (Vaswani et al., 2017) with 4 layers, 8 attention heads and a hidden size of 768. We use accuracy and F1 Score to evaluate prediction performances. Other details are presented in the Appendix.

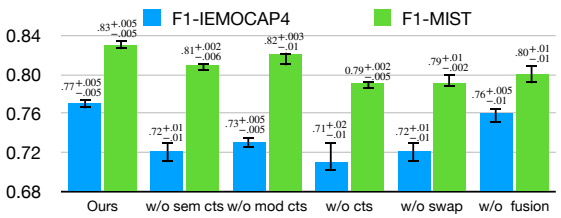


Figure 3: Ablation study on two benchmark datasets. Difference between different models is significant. Numerical results are presented in the appendix.

6 Results

From the results presented in Table 2 for the two benchmark datasets, we observe that the proposed framework significantly enhances performance on the IEMOCAP4 dataset, surpassing unimodal models. It improves the F1 score by over 6 points compared to the text-based RoBERTa model. It also outperforms the speech-based wav2vec2 model by over 10 points in F1 score. Similarly, on the MIST dataset, our method increases the F1 score by 5 points over the RoBERTa model and 10 points over the wav2vec2 model.

Furthermore, our proposed framework demonstrates superior performance over the other multimodal models (baselines). This is evident in the IEMOCAP4 dataset, where our method surpasses the top-performing baseline models (LRMP and LFMIM) by more than 3 points in both accuracy and F1 score. A similar trend is observed in the MIST dataset. These results ascertain the effectiveness of our method, highlighting its advantages over both unimodal and multimodal pre-trained models. To further evaluate our method on other SLU tasks, we also perform experiments on intent classification using SLURP dataset. We compare our method with the baseline of Roberta, Wav2vec2 and LFMIM which has the best overall performance on previous two tasks. The results are presented in Table 4. Our framework also achieved best performance on intent classification.

7 Analysis

7.1 Ablation Study

In order to evaluate the impact of various modules in our framework, we conduct ablation studies on the two benchmark datasets. The results are summarized in Figure 3.

Methods	IEMOCAP4-ASR	IEMOCAP4-TTS	MIST-ASR	MIST-TTS
LMF	0.73	0.69	0.80	0.80
MISA	0.74	0.68	0.80	0.80
LFMIM	0.74	0.70	0.80	0.81
LRMP	0.74	0.69	0.79	0.80
Ours	**0.77**	**0.74**	**0.82**	**0.83**

Table 5: Performance (F1 score) of different methods on two benchmark datasets in a unimodal scenario (modality-scarce scenario, when only one modality is available). **ASR** represents the results based on original speech data and generated text data. **TTS** refers to the results based on original text and generated speech.

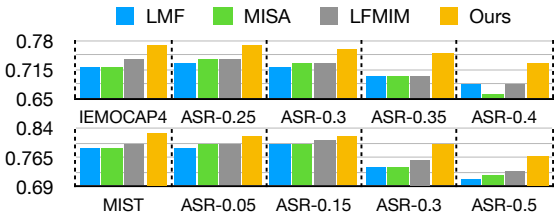


Figure 4: Performance of different methods on the IEMOCAP4 dataset and MIST dataset in a noisy text scenario. Numerical results are in the appendix.

Impact of Disentanglement. Our multi-level contrastive learning disentanglement is validated by separately removing each level of contrastive loss. The performance decrease underscores their necessity. Without either modality level ($\mathcal{L}_{\text{cts}}^{\text{sem}}$) or semantic level contrastive loss ($\mathcal{L}_{\text{cts}}^{\text{mod}}$), the model focuses only on semantic-/modality-related information, leading to subpar performance. Interestingly, the performance degradation is more severe without semantic level contrastive loss, indicating its greater importance. When both levels of contrastive loss are omitted, the performance significantly deteriorates, reaffirming the value of our proposed multi-level contrastive learning.

Impact of Swap Reconstruction. Without our proposed swap reconstruction mechanism, there is a degradation of over 4 points in both accuracy and F1 score. This suggests that our mechanism enhances the quality of modality-invariant representation by improving alignment between modality-invariant representations from different modalities.

Impact of Multi-Head Fusion. To underscore the significance of our multi-head fusion, we substitute it with a traditional attention-based fusion mechanism proposed in (Hazarika et al., 2020). Results reveal a performance drop of 1 point in F1 score on IEMOCAP4, and over 2 points on MIST. Thus, our mechanism improves representation quality.

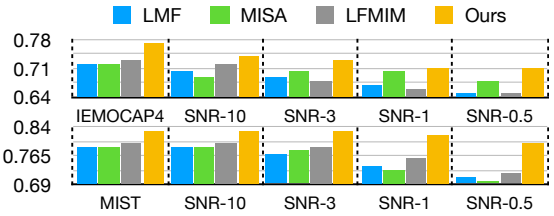
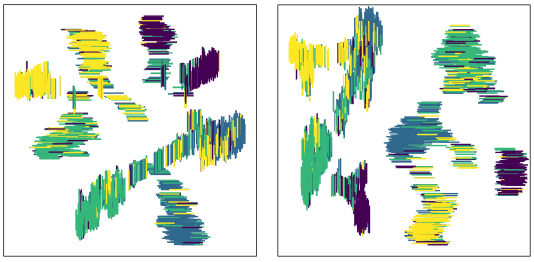


Figure 5: Performance of different methods on the IEMOCAP4 dataset and MIST dataset in a noisy speech scenario. Numerical results are in the appendix.



(a) modality-invariant (b) modality-specific

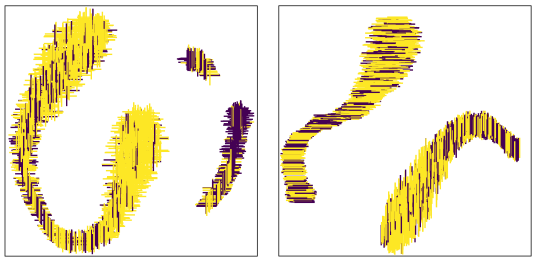
Figure 6: t-SNE visualization on IEMOCAP4. The horizontal line ('-') denotes the representations extracted from speech and vertical line ('|') denotes the representations extracted from text. Lines in purple denote the emotion Angry, lines in blue denote Happy, lines in green denote Neutral and lines in yellow denote Sad.

7.2 Qualitative Analysis

Integrating modalities. We present a qualitative analysis to demonstrate the effectiveness of integrating different modalities. Table 3 shows that on the IEMOCAP4 dataset, predictions based on a single modality may be inaccurate. However, when both speech and text are incorporated, the predictions are correct. This pattern is consistent with the results from the MIST dataset.

Disentangled representations. We note that the modality-invariant and modality-specific representations capture intended information as shown in the visualizations for IEMOCAP4 and MIST datasets (see Figures 6 and 7). Importantly, we note that the modality-invariant representations conform to the input and do not degenerate by posterior collapse (Lucas et al., 2019; Wang et al., 2021).

Visualization. Based on our multi-level contrastive learning, the modality-related information is expected to be encoded into modality-specific representations and semantic-related information is expected to be encoded into modality-invariant representations. Figures 6 and 7, visualize the IEMOCAP4 contextual embeddings and MIST contextual embeddings for the respective test set sentences. From Figure 6 we observe that: (1) the modality-



(a) modality-invariant (b) modality-specific

Figure 7: t-SNE visualization on MIST. The horizontal line ('-') denotes the representations extracted from speech and vertical line ('|') denotes the representations extracted from text. Lines in purple denote the figurative usage and lines in yellow denote the literal usage.

specific representations are well separated into the spoken and written clusters demonstrating the correspondence to specific modalities, (2) the modality-invariant representations capture the semantic content of the emotions and are broadly clustered corresponding to the input distribution. It is also noteworthy that the Neutral emotion (green lines) and the Happy emotion (blue lines) are difficult for the model to distinguish, which is in alignment with the observations noted by Busso et al. (2008). Likewise, most of the wrongly annotated Neutral examples are annotated as Happiness and Excitement.(3) The modality-specific representations extracted from both modalities cluster into four groups that roughly correspond to the four emotions. Additionally, we note that there are emotions that span both modalities (e.g., Angry denoted by purple). Therefore, the emotion-related features are disentangled into their modality-specific representations for the emotion recognition task. Similarly, in Figure 7, the modality-specific representations are well separated whereas the modality-invariant representations are overlapped (and conform to the input distribution). Note that the yellow lines are much more than the purple lines, which is in alignment with the statistics in Table 2 where the literal examples are more numerous than the figurative examples. Besides, we observe that the modality-invariant representations cluster into two groups which correspond to the two usages. Here again, we note that the task-related features are well captured by the modality-invariant representations for the idiom usage recognition task.

7.3 Efficacy in Modality-Scarce Scenarios

Parallel multimodal data, crucial for pre-training and fine-tuning, may not always be available. Unlike previous studies, we explore our framework in a unimodal scenario, by generating data for the missing modality.For example, we use ASR (Whisper model (Radford et al., 2023)) to generate text from speech data and a TTS model (Bark model[1]) to generate speech from text data. CRISP, as shown in Table 5, outperforms all baselines across both benchmark datasets, even in an unimodal scenario. Additionally, we observe that the results vary based on the modality of the generated data and the specific benchmark dataset. On IEMOCAP4, the results using original speech and generated text data are the best, suggesting that multimodal emotion recognition relies more on speech; and generated speech may lack crucial features like tone and stress. However, on MIST, results using original text and generated speech data are better, indicating a greater reliance on text for this task.

7.4 Efficacy in the Presence of Noise

Our framework's robustness against noise in data is analyzed against baseline methods with noise levels varying across the different modalities. For text data, we employ various ASR models (Whisper-tiny, small, base, and large (Radford et al., 2023)) to generate text from original speech, with different word error rates (WER) indicating the noise levels. For speech data, we introduce varying levels of white noise into the original speech signals, with the noise level represented by the Signal-to-Noise Ratio (SNR; lower SNR indicates higher noise). As Figures 4 and 5 illustrate, our model is robust to both text and speech noise.

8 Conclusion and Future Work

We introduced a novel approach, CLASP, that uses unimodal pre-trained models for multimodal tasks, eliminating the need for parallel speech-text corpora during pre-training. Our framework includes a disentanglement mechanism, a swap reconstruction mechanism, and a multimodal representation fusion module. Empirical evaluations on the SLU and NLU tasks justify the efficacy of CLASP and its broad applicability. Notably, it demonstrated robustness against noisy multimodal data, highlighting its real-world utility and adaptability. Additionally, given the substantial size of CLASP, future efforts could focus on developing parameter-efficient methods for its training.

[1]https://github.com/suno-ai/bark

9 Limitations

One drawback of our proposed framework lies in the large size of our proposed framework. Due to the fact we utilize two unimodal pre-trained models as text encoder and speech encoder, the size of the whole framework is very large. Future works could be done to propose parameter-efficient methods for training of our proposed whole framework.

Furthermore, compared to a jointly pre-trained speech-text model, our proposed method requires two uni-modal pre-trained models, which is still inconvenient. Actually our proposed method could be utilized for jointly pre-training a speech-text model, which we leave for future work.

Another limitation lies in our dataset that was automatically labelled resulting in a noisy training data. Future works could improve the quality and scale of our proposed dataset for multimodal idiom usage detection.

Besides, our method still requires a small number of parallel speech and text data for downstream fine-tuning and a large number of unimodal data for pre-training of the text encoder and speech encoder. However, for some languages and dialects, especially the spoken ones, parallel speech-text data and large-scale unimodal pre-training data might not be available. Therefore, it is necessary to explore the scenarios where no parallel speech-text data and large-scale unimodal pre-training data are available, which we leave for future work.

Acknowledgements

This research was supported by the National Science Foundation under Grant No. IIS 2230817 and in part by the U.S. National Science Foundation and Institute of Education Sciences under Grant No. 2229612.

References

Ankur Bapna, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang. 2021. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. arXiv preprint arXiv:2110.10329.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7252–7262.

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, 42:335–359.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro Moreno, Ankur Bapna, and Heiga Zen. 2022. Maestro: Matched speech text representations through modality matching. arXiv preprint arXiv:2204.03409.

Yung-Sung Chuang, Chi-Liang Liu, Hung-yi Lee, and Lin-shan Lee. 2020. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021. Improving multimodal fusion via mutual dependency maximisation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 231–245.

Susan R Fussell and Mallie M Moss. 2014. Figurative language in emotional communication. In Social and cognitive approaches to interpersonal communication, pages 113–141. Psychology Press.

Raymond W Gibbs Jr, John S Leggitt, and Elizabeth A Turner. 2002. What's special about figurative language in emotional communication? In The verbal communication of emotions, pages 133–158. Psychology Press.

Sam Glucksberg and Matthew S McGlone. 2001. Understanding figurative language: From metaphor to idioms. 36. Oxford University Press.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 279–287.

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2019. Casting a wide net: robust extraction of potentially idiomatic expressions. arXiv preprint arXiv:1911.08829.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9180–9192.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM international conference on multimedia, pages 1122–1131.

Yu Kang, Tianqiao Liu, Hang Li, Yang Hao, and Wenbiao Ding. 2022. Self-supervised audio-and-text pre-training with extremely low-resource parallel data. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 10875–10883.

Minjeong Kim, Gyuwan Kim, Sang-Woo Lee, and Jung-Woo Ha. 2021. St-bert: Cross-modal language model pre-training for end-to-end spoken language understanding. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7478–7482. IEEE.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Hang Li, Wenbiao Ding, Yu Kang, Tianqiao Liu, Zhongqin Wu, and Zitao Liu. 2021. Ctal: Pre-training cross-modal transformer for audio-and-language representations. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3966–3977.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning, pages 12888–12900. PMLR.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022b. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10965–10975.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5491–5496.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2247–2256.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. 2019. Understanding posterior collapse in generative latent variable models.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5206–5210. IEEE.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning, pages 28492–28518. PMLR.

Vin Sachidananda, Shao-Yen Tseng, Erik Marchi, Sachin Kajarekar, and Panayiotis Georgiou. 2022. Calm: Contrastive aligned audio-language multirate and multimodal representations. arXiv preprint arXiv:2202.03587.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In International conference on intelligent text processing and computational linguistics, pages 1–15. Springer.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. Interspeech 2019.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 658–670.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, et al. 2022. Unified speech-text pre-training for speech translation and recognition. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1488–1499.

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.

Yixin Wang, David Blei, and John P Cunningham. 2021. Posterior collapse and latent variable non-identifiability. Advances in Neural Information Processing Systems, 34:5443–5455.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In Proceedings of the 30th ACM International Conference on Multimedia, pages 1642–1651.

Tianshu Yu, Haoyu Gao, Ting-En Lin, Min Yang, Yuchuan Wu, Wentao Ma, Chao Wang, Fei Huang, and Yongbin Li. 2023. Speech-text pre-training for spoken dialog understanding with explicit cross-modal alignment. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7900–7913.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1103–1114.

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

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

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. Transactions of the Association for Computational Linguistics, 9:1546–1562.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In Proceedings of the 30th ACM International Conference on Multimedia, pages 1688–1697.

Jianing Zhou, Ziheng Zeng, and Suma Bhat. 2023. Clcl: Non-compositional expression detection with contrastive learning and curriculum learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 730–743.

Methods	exShort	Short	Medium	Long
Text Only	0.55	0.68	0.68	0.77
Audio Only	0.67	0.62	0.64	0.70
Ours	0.69	0.73	0.72	0.84

Table 6: Performance on examples with different lengths.

Methods	Happy	Sad	Angry	Neutral
Text Only	0.75	0.71	0.70	0.67
Audio Only	0.64	0.71	0.76	0.72
Ours	0.78	0.78	0.79	0.73

Table 7: Performance on examples with different emotions.

A Experimental Settings

The model is trained with a batch size of 4 for 25 epochs. Adam optimizer (Kingma and Ba, 2014) is used and the learning rate is set to 1×10^{-5}. All the other parameters are set to their default. The hyperparameters of α, β γ and ζ are set to 0.01, 0.01, 0.01 and 0.05 respectively. All of our experiments were conducted using two GPUs with 16GB RAM (NVIDIA V100).

We utilize grid search to tune these hyperparameters. Each hyperparameter is selected from 0.01, 0.05, 0.1, 0.5. Our model achieves the best performance for α, β and γ each equal to 0.01 and δ = 0.05. For other values of the hyperparameters, our model performance degrades at most by 0.02 on IEMOCAP4. We only perform the hyperparameter tuning on IEMOCAP4 and fix the optimal hyperparameters for MIST.

B Analysis

Performance with respect to length of utterance based on IEMOCAP4: Considering that the amount of information at the utterance level can impact emotion recognition, we calculate the accuracy on extremely short examples (only 1 word), short examples (less than 4 words but more than 1 word), medium examples (between 4 and 10 words), and long examples (more than 10 words). The results are presented in Table 6. Compared with the unimodal models, we notice the gains in performance on account of harnessing information in both modalities across all length categories.

We also report the accuracy on the emotion categories of Happy, Sad, Angry and Neutral in Table 7. We observe that our proposed model that effectively leverages both audio and text outperforms unimodal models in all the categories, further demonstrating the benefits of multimodal language understanding.

Methods	IEMOCAP4		MIST	
	Acc	F1	Acc	F1
RoBERTa-Text Only	0.71	0.7	0.78	0.77
wav2vec2-Audio Only	0.67	0.67	0.73	0.73
LMF	0.72	0.71	0.79	0.78
MISA	0.72	0.71	0.79	0.78
LFMIM	0.74	0.73	0.8	0.8
LRMP	0.74	0.74	0.8	0.79
Ours	**0.77**	**0.77**	**0.83**	**0.83**

Table 8: Performance of different methods on two benchmark datasets. Best performance is labeled in bold.

Methods	IEMOCAP4		MIST	
	Acc	F1	Acc	F1
Ours	**0.77**	**0.77**	**0.83**	**0.83**
w/o $\mathcal{L}_{cts}^{\text{sem}}$	0.73	0.72	0.81	0.81
w/o $\mathcal{L}_{cts}^{\text{mod}}$	0.74	0.73	0.82	0.82
w/o $\mathcal{L}_{cts}^{\text{sem}}$ and $\mathcal{L}_{cts}^{\text{mod}}$	0.72	0.71	0.79	0.79
w/o $\mathcal{L}_{recon}$	0.73	0.72	0.8	0.79
w/o multi-head fusion	0.76	0.76	0.81	0.8

Table 9: Ablation study on two benchmark datasets. Best performance is labeled in bold.

Methods	IEMOCAP4	ASR-0.25	ASR-0.3	ASR-0.35	ASR-0.4
LMF	0.72	0.73	0.72	0.70	0.68
MISA	0.72	0.74	0.73	0.70	0.66
LFMIM	0.74	0.74	0.73	0.70	0.68
Ours	**0.77**	**0.77**	**0.76**	**0.75**	**73**

Table 10: Performance (F1 score) of different methods on IEMOCAP4 under the noisy scenario. Different numbers after ASR refers to different WER. Best performance is labeled in bold.

Methods	IEMOCAP4	SNR-10	SNR-3	SNR-1	SNR-0.5
LMF	0.72	0.70	0.69	0.67	0.65
MISA	0.72	0.69	0.70	0.70	0.68
LFMIM	0.73	0.72	0.68	0.66	0.65
Ours	**0.77**	**0.74**	**0.73**	**0.71**	**0.71**

Table 11: Performance (F1 score) of different methods on IEMOCAP4 under the noisy scenario. The numbers after SNR refer to different levels of white noise.

C Case Study

Here we elaborate further on the qualitative analysis as shown in Table 14. We observe that on the IEMOCAP4 dataset, our proposed method effectively incorporates the speech and text modality,

Methods	MIST	ASR-0.05	ASR-0.15	ASR-0.3	ASR-0.5
LMF	0.79	0.79	0.80	0.74	0.71
MISA	0.79	0.80	0.80	0.74	0.72
LFMIM	0.80	0.80	0.81	0.76	0.73
Ours	**0.83**	**0.82**	**0.82**	**0.80**	**0.77**

Table 12: Performance of different methods on the MIST dataset in a noisy text scenario.

Methods	MIST	SNR-10	SNR-3	SNR-1	SNR-0.5
LMF	0.79	0.79	0.77	0.74	0.71
MISA	0.79	0.79	0.78	0.73	0.70
LFMIM	0.80	0.80	0.79	0.76	0.72
Ours	**0.83**	**0.83**	**0.83**	**0.82**	**0.80**

Table 13: Performance of different methods on MIST in a noisy speech scenario.

which is especially noteworthy on short utterances, such as 'Uh-huh.' or 'Right', where the use of a single modality can be insufficient.

Task	Given Example	Truth	Text	Speech	Ours
IEMOCAP4	Uh-huh.	Happy	Sad	Neutral	Happy
IEMOCAP4	Yeah so- That –that did go through.	Neutral	Sad	Happy	Neutral
IEMOCAP4	Right.	Happy	Sad	Angry	Happy
IEMOCAP4	You can't make fun of me because I'm not going to school, though, because I'm going to be working though.	Happy	Neutral	Neutral	Happy
IEMOCAP4	He was out here when it broke.	Sad	Neutral	Happy	Sad
MIST	his book of leaves would not have told him in my own handwriting that i believed in his better nature	Figurative	Literal	Literal	Figurative
MIST	fortunately he had no children to run the risk of madness in their turn	Figurative	Literal	Literal	Figurative
MIST	which made the animal's legs almost give waydartanian burst out laughing as he said take care o planchet	Figurative	Literal	Literal	Figurative
MIST	homo turned his head now and then to make sure that guenplane was behind him	Literal	Figurative	Figurative	Literal
MIST	in a moment it came again a thumping of the old knocker on the front-inner door	Literal	Figurative	Figurative	Literal

Table 14: Qualitative Examples