Unsupervised Multimodal Clustering for Semantics Discovery in Multimodal Utterances

Hanlei Zhang[1], Hua Xu[1]*, **Fei Long[1], Xin Wang[1,2,3], Kai Gao[2]**

[1]State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University,
[2]School of Information Science and Engineering, Hebei University of Science and Technology
[3]Samton (Jiangxi) Technology Development Co.,Ltd, Nanchang 330036, China
zhang-hl20@mails.tsinghua.edu.cn, xuhua@tsinghua.edu.cn

Abstract

Discovering the semantics of multimodal utterances is essential for understanding human language and enhancing human-machine interactions. Existing methods manifest limitations in leveraging nonverbal information for discerning complex semantics in unsupervised scenarios. This paper introduces a novel unsupervised multimodal clustering method (UMC), making a pioneering contribution to this field. UMC introduces a unique approach to constructing augmentation views for multimodal data, which are then used to perform pre-training to establish well-initialized representations for subsequent clustering. An innovative strategy is proposed to dynamically select high-quality samples as guidance for representation learning, gauged by the density of each sample's nearest neighbors. Besides, it is equipped to automatically determine the optimal value for the top-K parameter in each cluster to refine sample selection. Finally, both high- and low-quality samples are used to learn representations conducive to effective clustering. We build baselines on benchmark multimodal intent and dialogue act datasets. UMC shows remarkable improvements of 2-6% scores in clustering metrics over state-of-the-art methods, marking the first successful endeavor in this domain. The complete code and data are available at https://github.com/thuiar/UMC.

1 Introduction

Discovering the semantics of dialogue utterances in unsupervised multimodal data requires integrating various modalities (i.e., text, video, and audio) to effectively mine the complicated semantics inherent in multimodal language. Conventional methods for semantics discovery typically focus solely on the text modality with clustering algorithms (Zhang et al., 2021a, 2023), failing to leverage the rich multimodal information in the real world (e.g., body language, facial expressions, and tones).

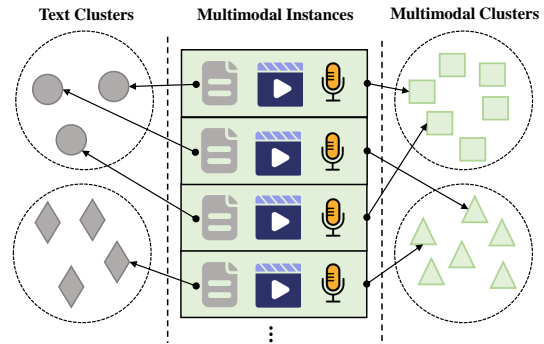
* Hua Xu is the corresponding author.

Figure 1: Text-only clustering deviates from real multimodal utterance semantics, highlighting the need of multimodal information in semantics discovery.

However, we argue that non-verbal modalities (i.e., video and audio) also play a critical role when performing unsupervised clustering. Taking Figure 1 as an example, relying solely on textual information yields clustering results that differ from the ground truth of multimodal cluster allocations (a detailed analysis on real-world examples is available in Appendix A), suggesting that non-verbal modalities can provide useful cues for semantics discovery. Moreover, effectively capturing multimodal interactions can yield more powerful and robust representations, thereby better addressing the challenges of ambiguous intent-cluster boundaries found in text-based clustering (see Section 6.3 and Appendix J). Discovering multimodal utterance semantics holds significant promise for a variety of applications, including video content recommendation, efficient multimodal data annotation, and virtual human technologies (detailed in Appendix B).

Understanding semantics in multimodal utterances has attracted much attention with the boom in multimodal language analysis (Poria et al., 2019; Saha et al., 2021b; Zhang et al., 2022a). For example, Saha et al. (2021b) annotated multimodal dialogue act (DA) labels on two popular multimodal multi-party conversational datasets (Busso

et al., 2008; Poria et al., 2019) and performed DA recognition using attention sub-networks build upon modality encoders. Zhang et al. (2022a) pioneered multimodal intent analysis, introducing a new dataset with multimodal intent labels and establishing baselines with three multimodal fusion methods (Tsai et al., 2019; Rahman et al., 2020; Hazarika et al., 2020). However, these works remain restricted within supervised tasks, i.e., the training target for each piece of data is known, which is not applicable in unsupervised scenarios.

In contrast, semantics discovery is an emerging field in NLP. It fundamentally operates as a clustering task and has seen the development of many unsupervised (Cheung and Li, 2012; Padmasundari and Bangalore, 2018; Haponchyk et al., 2018; Zhang et al., 2023) and semi-supervised (Lin et al., 2020; Zhang et al., 2021c, 2022b; Zhou et al., 2023) methods. However, these methods are primarily designed for the text-only modality and lack proficiency in handling the diverse modalities encountered in real-world scenarios. Thus, there is a lack of multimodal clustering methods for discovering utterance semantics, posing two challenges: (1) determining how to leverage information from nonverbal modalities to complement the text modality in clustering and (2) devising ways to fully exploit multimodal unlabeled data to learn clustering-friendly representations.

To address these challenges, we introduce UMC, a novel unsupervised multimodal clustering algorithm for semantics discovery, as shown in Figure 2. We utilize the capabilities of the pre-trained language model (Devlin et al., 2019) to process text data. For the video and audio modalities, deep features are initially extracted using powerful backbones from computer vision and speech signal processing. Two transformer encoders are then employed to capture the deep semantics of these features. The text modality is designated as the anchor, guiding the learning of the other modalities. For this purpose, we concatenate features from all three modalities and mask the video or audio features with zero vectors, creating two sets of positive augmentation views. These multimodal representations and their augmentations are applied to an unsupervised contrastive loss, yielding well-initialized representations for subsequent process.

To fully mine the semantic similarities among unsupervised multimodal data, we introduce a novel strategy that initially selects high-quality samples.

This strategy employs a dynamic sample selection threshold t, aiming to select the highest-quality t percent of samples in each iteration for training. This selection is based on a unique mechanism that calculates the density of each sample within its respective cluster and ranks them accordingly. Besides, an evaluation process is designed to automatically determine the optimal parameters for the top-K nearest neighbors from a set of candidates. After selecting high-quality samples, we propose a sequential process for multimodal representation learning. This process begins by learning from high-quality samples using supervised contrastive loss and then refines the remaining low-quality samples using unsupervised contrastive loss. This two-step approach promotes beneficial intra-class and inter-class relations among high-quality samples while pushing apart low-quality samples, thereby generating representations conducive to clustering. The entire process is repeated until the sample selection threshold t is met.

We summarize our contributions as follows:

In this work, we make a pioneering contribution by formulating the challenging multimodal semantics discovery task. To solve this problem, we first introduce a novel method for constructing positive augmentations for multimodal data, effectively leveraging non-verbal modalities for unsupervised pre-training, which provides a good initialization for unsupervised clustering.

Then, we propose a new clustering algorithm, UMC, which features an innovative high-quality sample selection strategy and a sequential representation learning method between high- and low-quality samples, resulting in excellent performance across both single and multimodal modalities.

Finally, we establish baselines using benchmark multimodal intent and dialogue datasets. Extensive experiments show that the proposed UMC outperforms state-of-the-art clustering algorithms by a notable margin of 2-6% scores in standard clustering metrics. To the best of our knowledge, this is the first successful attempt at leveraging multiple modalities for unsupervised clustering, marking a substantial advancement in this area.

2 Related Works

2.1 Unsupervised Clustering

Unsupervised clustering is fundamental in machine learning. Classic clustering methods like K-Means (MacQueen et al., 1967) and Agglomerative
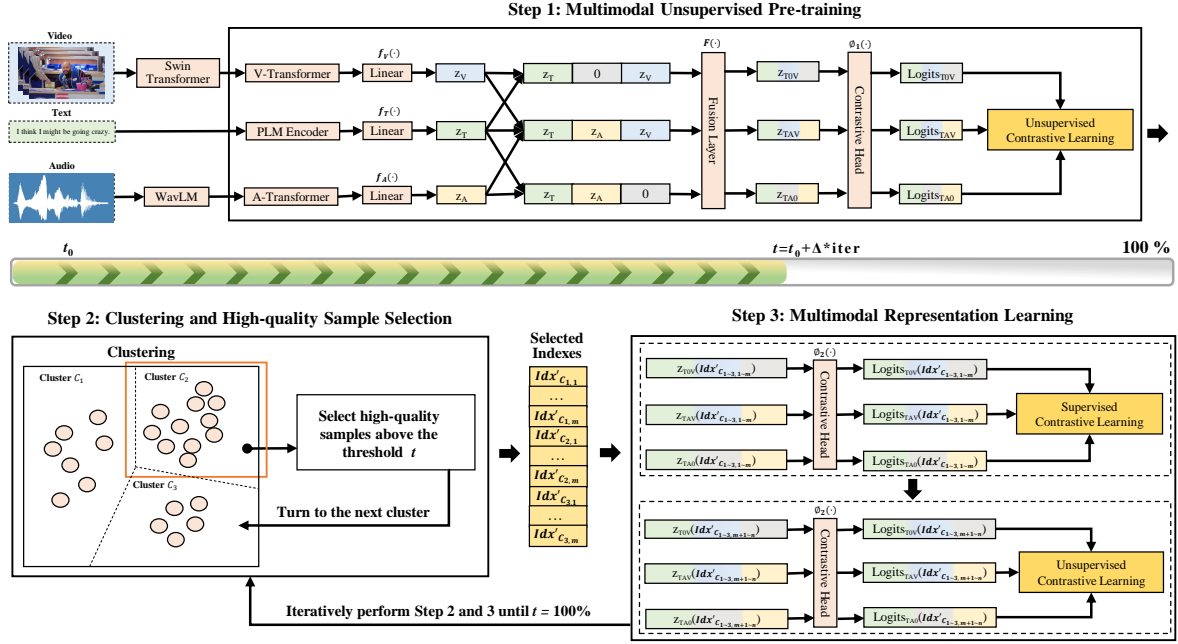
Figure 2: Overview of our proposed unsupervised multimodal clustering algorithm UMC.

Clustering (Gowda and Krishna, 1978) iteratively assign clusters until convergence based on features. Deep clustering methods, like DEC (Xie et al., 2016) and DCN (Yang et al., 2017), enhance this process by jointly clustering and feature learning, employing stacked autoencoders (Vincent et al., 2010). DeepCluster (Caron et al., 2018) uses cluster assignments as guidance for feature learning.

Recent methods using contrastive learning (Chen et al., 2020) have achieved the state-of-the-art performance. For instance, SCCL (Zhang et al., 2021a) combines instance-level contrastive learning with cluster refinement from target distributions. CC (Kumar et al., 2022) optimizes contrastive losses at both instance and cluster levels to generate clustering-friendly representations. However, these methods focus on merely the single text or image modality and fall short with multimodal data. MCN (Chen et al., 2021) is tailored for multimodal clustering, learning a unified representation from all modalities and applying cross-modal contrastive losses during clustering. However, MCN struggles with complex utterance semantics.

2.2 Intent Discovery

Intent discovery is a key challenge in NLP, with numerous clustering methods developed to address it. Early methods (Hakkani-Tür et al., 2015; Haponchyk et al., 2018) use weakly supervised signals to aid in clustering but struggle to capture the high-level semantics in text. Recent methods (Lin

et al., 2020; Zhang et al., 2021c; Mou et al., 2022; Zhang et al., 2022b; Mou et al., 2023; Zhou et al., 2023; Shi et al., 2023) exploit limited labeled data to guide the feature learning process for clustering.

However, these methods suffer a substantial decrease in performance in totally unsupervised scenarios. USNID (Zhang et al., 2023) proposes a novel centroid-guided mechanism with a pre-training strategy, achieving significant improvements over previous methods. Yet, USNID also falls short in handling multimodal data. See Appendix C for more related works on multi-view clustering and multimodal language analysis.

3 Problem Formulation

For the task of multimodal semantics discovery, we are provided with a multimodal intent or dialogue act dataset $\mathcal{D}_{\text{mm}} = \{(s_i^{\text{T}}, s_i^{\text{A}}, s_i^{\text{V}}) | y_i \in \mathcal{I}, i = 1, ..., N\}$, where each ith instance s_i contains multimodal utterances, including s_i^{T}, audio s_i^{A}, and video s_i^{V}. Here, N represents the total number of instances. The ground-truth label y_i, belonging to the set of intent or dialogue act classes $\mathcal{Y} = \{y_i\}_{i=1}^{K_\mathcal{Y}}$, remains unseen during training and validation and is only available during testing. The number of classes is denoted by $K_\mathcal{Y}$.

The objective is to learn a multimodal neural network $\mathcal{F}(\cdot)$ capable of obtaining multimodal representations z conducive to clustering. These representations are subsequently employed to divide

the set $\{\boldsymbol{s}_i\}_{i=1}^N$ into $K_{\mathcal{Y}}$ groups.

4 Methodologies

4.1 Multimodal Representation

To obtain multimodal representations, we first extract deep features from text, video, and audio modalities. For text, we employ the pre-trained language model (PLM), BERT (Devlin et al., 2019) as the encoder, fine-tuning it on the text inputs $\boldsymbol{s}^{\mathrm{T}}$. The initial [CLS] token embedding, $\boldsymbol{x}_{\mathrm{T}} \in \mathbb{R}^{D_{\mathrm{T}}}$, serves as the sentence-level representation, where D_{T} is the feature dimension of 768. We then incorporate a linear layer, represented as $f_{\mathrm{T}}(\cdot)$, yielding $\boldsymbol{z}_{\mathrm{T}} \in \mathbb{R}^{D_H}$. Here, H indicates a dimensionally reduced space, enhancing computational efficiency and accentuating primary features.

For non-verbal modalities, we use semantically rich features as inputs as suggested in (Saha et al., 2020; Zhang et al., 2022a). For video, we employ the Swin Transformer (Liu et al., 2021) to extract video feature representations $\boldsymbol{x}_{\mathrm{V}} \in \mathbb{R}^{L_{\mathrm{V}} \times D_{\mathrm{V}}}$ at the frame level from the video inputs $\boldsymbol{s}^{\mathrm{V}}$. Here, L_{V} represents the video length, and D_{V} is the feature dimension of 1024. For audio $\boldsymbol{s}^{\mathrm{A}}$, we first extract audio waveforms as in (Zhang et al., 2022a) and then use the WavLM (Chen et al., 2022) to obtain features $\boldsymbol{x}_{\mathrm{A}} \in \mathbb{R}^{L_{\mathrm{A}} \times D_{\mathrm{A}}}$. Here, L_{A} and D_{A} denote the audio length and feature dimension of 768, respectively. Unsupervised multimodal clustering can benefit from these two powerful non-verbal features extracted from the Swin Transformer and WavLM models. A comparison between them and other multimodal features is shown in Appendix D.

For both audio and video modalities, initially introduce a linear layer $f_M(\cdot)$ in alignment with the text modality. Subsequently, we apply the multi-headed attention mechanism with the Transformer (Vaswani et al., 2017) encoder, adeptly capturing intricate semantic relationships and temporal nuances. Eventually, in line with (Tsai et al., 2019), the last sequence elements are employed to derive the sentence-level representation $\boldsymbol{z}_M$:

$$\boldsymbol{z}_M = \mathrm{Transformer}(f_M(\boldsymbol{x}_M))[-1], \quad (1)$$

where $M \in \{\mathrm{A}, \mathrm{V}\}$, and $\boldsymbol{z}_M \in \mathbb{R}^{D_H}$.

Following this, we concatenate the representations $\boldsymbol{z}_{\mathrm{T}}$, $\boldsymbol{z}_{\mathrm{A}}$, and $\boldsymbol{z}_{\mathrm{V}}$ and pass them through a non-linear fusion layer, denoted as $\mathcal{F} : \mathbb{R}^{3D_H} \to \mathbb{R}^{D_H}$. This layer is designed to learn cross-modal interactions, yielding the combined representation

$\boldsymbol{z}_{\mathrm{TAV}} \in \mathbb{R}^{D_H}$:

$$\boldsymbol{z}_{\mathrm{TAV}} = \mathcal{F}(\mathrm{Concat}(\boldsymbol{z}_{\mathrm{T}}, \boldsymbol{z}_{\mathrm{A}}, \boldsymbol{z}_{\mathrm{V}})), \quad (2)$$

where $\mathcal{F}$ is defined as $W_1 \sigma_{\mathrm{GELU}}(\mathrm{Dropout}(\cdot)) + b_1$. Here, σ_{GELU} represents the GELU activation function, and W_1 and b_1 are the corresponding weight and bias matrices, respectively. Subsequently, we employ $\boldsymbol{z}_{\mathrm{TAV}}$ and its augmentations for further clustering and representation learning.

4.2 Multimodal Unsupervised Pre-training

Effective pre-training strategies can provide well-initialized representations conducive to clustering (Zhang et al., 2023). Unsupervised contrastive learning (Chen et al., 2020) has emerged as an effective approach for unsupervised clustering (Zhang et al., 2021a; Li et al., 2021). It pushes apart samples and makes them distribute uniformly in the feature space while capturing implicit similarity relations between augmentations. However, existing methods often fall short in providing effective augmentations for multimodal data. In this work, we introduce a novel method of non-verbal modality masking to address this gap.

Given the predominant role of the text modality in intent analysis, we retain it as the core modality and mask either the video or audio modality for data augmentation. For the ith sample $\boldsymbol{z}_{\mathrm{TAV},i}$ in a minibatch of B samples, either the video or audio modality is replaced with zero vectors. Eq. 2 is used to derive $\boldsymbol{z}_{\mathrm{TA0},i}$ and $\boldsymbol{z}_{\mathrm{T0V},i}$ as positively augmented samples. For each positive pair (i, j) among the generated $3B$ augmented samples, we apply the multimodal unsupervised contrastive learning loss:

$$\mathcal{L}_{i,j}^{\mathrm{mucl}} = \\ -\log \left(\frac{\exp(\mathrm{sim}(\phi_1(\boldsymbol{z}_i), \phi_1(\boldsymbol{z}_j))/\tau_1)}{\sum_k \mathbb{I}_{[k \neq i]} \exp(\mathrm{sim}(\phi_1(\boldsymbol{z}_i), \phi_1(\boldsymbol{z}_k))/\tau_1)} \right), \quad (3)$$

where $\boldsymbol{z}_i \in \{\boldsymbol{z}_{\mathrm{TAV},i}, \boldsymbol{z}_{\mathrm{TA0},i}, \boldsymbol{z}_{\mathrm{T0V},i}\}$, $\mathrm{sim}(\cdot)$ refers to the dot product operation on two L2-normalized vectors, and $\phi_1(\cdot)$ is a non-linear layer with ReLU activation, serving as the contrastive head. The parameter τ_1 represents the temperature, and $\mathbb{I}_{[\cdot]}$ is the indicator function, outputting 1 if and only if $j = i$, and 0 otherwise.

By masking the video or audio modality with zero vectors, the model can focus on learning the implicit similarities in the shared modalities among
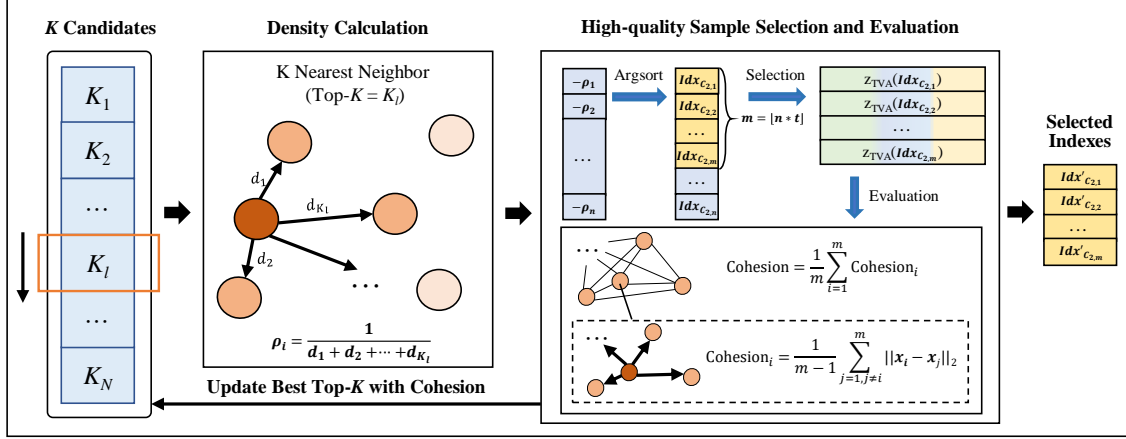
Figure 3: Pipeline of the high-quality sample selection mechanism.

positive pairs (i.e., text and video, text and audio, and text alone). This further encourages the model to capture intricate relationships and leverage complementary information across modalities.

4.3 Clustering and High-Quality Sample Selection

After pre-training, we employ the representations z_{TAV} to perform clustering. Specifically, we adopt the K-Means++ algorithm (Arthur and Vassilvitskii, 2007) for this task due to its advanced initial centroid selection technique that improves convergence over standard K-Means.

However, we observe that the cluster assignments obtained directly from K-Means++ are insufficiently high-quality to guide the learning of multimodal representations. To address this, we introduce a strategy to incrementally incorporate high-quality samples into the learning process. This is achieved through a curriculum-based method, where we progressively adjust the sample selection threshold, t, dictating the proportion of selected samples from each cluster for a given training iteration. The threshold t is linearly updated as follows:

$$t = t_0 + \Delta \cdot \text{iter}, \qquad (4)$$

where $t, t_0 \in [0, 1]$, t_0 is the initial threshold set to 0.1 (see Appendix E for a detailed discussion), iter is the iteration index within the epoch, and Δ is a preset positive increment, applied after each epoch.

To further refine clustering performance, we incorporate the centroid inheritance strategy as proposed in (Zhang et al., 2023). Specifically, K-Means++ is utilized only during the first training iteration. In subsequent iterations, the cluster centroids from the previous iteration are inherited as

initial centroids. This approach effectively leverages historical clustering information to guide and improve current clustering results.

Then, we need to identify high-quality samples for representation learning. We introduce a novel mechanism for selecting high-quality samples, as depicted in Figure 3. This mechanism comprises two main steps: density calculation and high-quality sample selection and evaluation.

4.3.1 Density Calculation

To discern high-quality samples within each cluster, we propose using density as the criterion. The underlying intuition is that high-quality samples are likely to exhibit high local density, whereas low-quality, anomalous, or falsely clustered data are expected to have low local density. For the ith sample, we compute its density, ρ_i, as the reciprocal of the average distance between $z_{\text{TAV},i}$ and its top-K nearest neighbors:

$$\rho_i = \frac{K_{\text{near}}}{\sum_{j=1}^{K_{\text{near}}} d_{ij}}, \qquad (5)$$

where K_{near} denotes the number of top-K nearest neighbors. d_{ij} represents the Euclidean distance between the ith sample and its jth nearest neighbor.

4.3.2 High-Quality Sample Selection and Evaluation

After calculating the density of each sample in each cluster, we rank them based on their densities in descending order. Specifically, for each sample in the kth cluster C_k with a density of ρ_i, we compute a sorted index list Idx_{C_k} as follows:

$$Idx_{C_k} = \text{argsort}(-[\rho_1, \rho_2, \cdots, \rho_n]), \qquad (6)$$

where argsort yields the indices that sort the densities in ascending order of the negative values, and n represents the number of samples in C_k. The high-quality samples are selected based on the highest densities. The number of selected highest-density samples has a proportion in cluster C_k above the threshold t. Let $m = \lfloor n * t \rfloor$, the chosen samples are denoted as: $z_{\text{TAV}}(Idx_{C_{k,1}}), \ldots, z_{\text{TAV}}(Idx_{C_{k,m}})$. $z_{\text{TAV}}(Idx_{C_{k,i}})$ is the ith selected sample feature in cluster C_k, based on the ordered density indices.

Considering that real-world data might not exhibit a uniform distribution across each class, assigning a fixed K_{near} to every cluster could compromise the precision of density calculations, subsequently affecting the selection of high-quality samples. To address this, we introduce an innovative method to automatically select the optimal K_{near}^k for each cluster C_k. Initially, we provide a candidate set $\{K_{\text{near},q}^k\}_{q=1}^u$, uniformly sampled based on the cluster size $|C_k|$. Specifically, $K_{\text{near},q}^k$ is defined as:

$$K_{\text{near},q}^k = \lfloor |C_k| \cdot (L + \Delta' \cdot (q-1)) \rfloor, \quad (7)$$

where L is the lower proportion bound with the constraint of $0 \leq L \leq 1$, Δ' is a fixed interval, and u is the number of candidates. Then, for each candidate $K_{\text{near},q}^k$, we use Eq. 6 to compute sorted indices $Idx_{C_k}^q$ and select a subset C_k^q with top-m samples. The quality of C_k^q is gauged through the cluster cohesion metric, measuring intra-cluster similarity. The cohesion of C_k^q is defined as:

$$\text{coh}(C_k^q) = \sum_{i=1}^m \text{coh}(C_{k,i}^q), \quad (8)$$

$$\text{coh}(C_{k,i}^q) = \quad (9)$$
$$\frac{1}{m-1} \sum_{j=1, j \neq i}^m d(z_{\text{TAV}}(Idx_{C_{k,i}^q}), z_{\text{TAV}}(Idx_{C_{k,j}^q})),$$

where m is the previously defined number of chosen samples, $d(\cdot)$ represents the Euclidean distance. The cohesion score can effectively capture the feature compactness and reflect the cluster quality. The optimal selected candidate index q_{opt} is calculated as:

$$q_{\text{opt}} = \underset{q}{\text{argmax}} \left\{ \text{coh}(C_k^q) \right\}. \quad (10)$$

That is, the optimal $K_{\text{near},q_{\text{opt}}}^k$ is selected by the candidate with the highest cluster cohesion score.

Subsequently, we use $\{K_{\text{near},q_{\text{opt}}}^k\}_{k=1}^{K_{\mathcal{Y}}}$ to obtain the selected high-quality indices $Idx' = \{Idx_{C_k}^{q_{\text{opt}}}\}_{k=1}^{K_{\mathcal{Y}}}$ with Eq. 5 and 6. These indices are then employed to select high-quality samples for subsequent representation learning.

4.4 Multimodal Representation Learning

The high-quality samples identified by the selected indices Idx' tend to have more reliable pseudo-labels, so we employ them as a guiding set to facilitate the learning of friendly representations for clustering. We aim to leverage these samples to capture high-level similarity relations between pairwise samples. To achieve this, we introduce the multimodal supervised contrastive loss:

$$\mathcal{L}_i^{\text{mscl}} =$$
$$\frac{-1}{|P(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\text{sim}(l_i, l_p)/\tau_2)}{\sum_j \mathbb{I}_{[j \neq i]} \exp(\text{sim}(l_i, l_j)/\tau_2)},$$
$$(11)$$

where $l_i = \phi_2(z_i)$, and ϕ_2 is a non-linear layer with ReLU activation, consistent with Eq. 3. Here, we perform the same data augmentation techniques as in section 4.2, and $l_i \in \{l_{\text{TAV},i}, l_{\text{T0V},i}, l_{\text{TA0},i}\}$. τ_2 denotes the temperature parameter. $P(i)$ is the set of indices for the augmented samples that share the same classes with l_i. With this loss, each sample can learn not only from its respective augmentations but also learn from the clustering information derived from high-quality pseudo-labels.

Conversely, low-quality samples are prone to erroneous clustering, where dissimilar samples may be grouped into the same class. This misgrouping can disrupt the integrity of the clustering process. To mitigate this issue, we propose the application of an unsupervised contrastive loss to these samples. This loss function is designed to increase the separation between distinct low-quality samples, thereby encouraging a more uniform distribution in the feature space, as supported by (Zhang et al., 2021a). Specifically, we use Eq. 3, replacing ϕ_1 with ϕ_3, and apply this modified equation to the remaining samples in the training set, excluding those with selected indices Idx'.

In our approach, we sequentially apply multimodal supervised contrastive learning to high-quality samples and unsupervised contrastive learning to low-quality samples. This two-step strategy is crafted to concurrently enhance multimodal representation learning and clustering process. The

Datasets	#C	#U	#Train	#Test
MIntRec	20	2,224	1,779	445
MELD-DA	12	9,988	7,990	1,998
IEMOCAP-DA	12	9,416	7,532	1,884

Table 1: Statistics of MIntRec, MELD-DA, IEMOCAP-DA datasets. # indicates the total number of sentences. #C and #U denote the number of classes and utterances.

training phase concludes when the sample selection threshold t (as defined in Eq. 4) reaches 100%. During the inference stage, we utilize the well-trained model to extract z_{TAV} and subsequently employ the K-Means++ algorithm for prediction.

5 Experiments

5.1 Datasets

We use MIntRec, MELD-DA, and IEMOCAP-DA as benchmark datasets for the multimodal semantics discovery task. The rationale for using these datasets is that the defined intents or dialogue acts typically exhibit a variety of distinct sentence-level semantics and possess properties of uncertainty in the open world, making them suitable for discovery in unsupervised scenarios. Detailed statistics of the three datasets are presented in Table 1, with further information on dataset specifics and their splits available in Appendix F.

5.2 Baselines

We compare UMC with the state-of-the-art unsupervised clustering methods from both NLP and CV, as well as multimodal clustering methods. The TEXTOIR platform (Zhang et al., 2021b) is used to reproduce the methods in NLP. Detailed descriptions of the baselines are follows:

SCCL (Zhang et al., 2021a): It jointly optimizes clustering and instance-level contrastive learning losses. The learning rate is set to 3e-5.

CC (Li et al., 2021): It employs dual non-linear heads to independently optimize instance-level and cluster-level contrastive learning losses. The learning rate is set to 3e-5.

USNID (Zhang et al., 2023): It performs strong data augmentation by randomly erasing words in a sentence. It also introduces a centroid-guided clustering mechanism to construct high-quality supervised signals for representation learning .

UMC (Text): This UMC variant excludes video and audio modalities during clustering. Unlike UMC, which uses multimodal augmentations, here

	Methods	NMI	ARI	ACC	FMI	Avg.
MIntRec	SCCL	45.33	14.60	36.86	24.89	30.42
	CC	47.45	22.04	41.57	26.91	34.49
	USNID	47.91	21.52	40.32	26.58	34.08
	MCN	18.24	1.70	16.76	10.32	11.76
	UMC (Text)	47.15	22.05	42.46	26.93	34.65
	UMC	**49.26**	**24.67**	**43.73**	**29.39**	**36.76**
M-DA	SCCL	22.42	14.48	32.09	27.51	24.13
	CC	23.03	13.53	25.13	24.86	21.64
	USNID	20.80	12.16	24.07	23.28	20.08
	MCN	8.34	1.57	18.10	15.31	10.83
	UMC (Text)	19.57	16.29	33.40	30.81	25.02
	UMC	**23.22**	**20.59**	**35.31**	**33.88**	**28.25**
I-DA	SCCL	21.90	10.90	26.80	24.14	20.94
	CC	23.59	12.99	25.86	24.42	21.72
	USNID	22.19	11.92	27.35	23.86	21.33
	MCN	8.12	1.81	16.16	14.34	10.11
	UMC (Text)	20.01	18.15	32.76	31.10	25.64
	UMC	**24.16**	**20.31**	**33.87**	**32.49**	**27.71**

Table 2: Results on MIntRec, MELD-DA (M-DA), and IEMOCAP-DA (I-DA) datasets.

we apply *dropout* twice to generate positive augmentations for contrastive learning.

MCN (Chen et al., 2021): It employs an online K-Means algorithm to dynamically determine cluster centers and periodically update them. However, we find its performance drops with online clustering. Thus, we modify it to perform K-Means on the full dataset to ensure optimal performance.

5.3 Evaluation Metrics

Following (Fahad et al., 2014; Saxena et al., 2017), we use four standard clustering metrics to evaluate the clustering performance, including Normalized Mutual Information (NMI), Accuracy (ACC), Adjusted Rand Index (ARI), and Fowlkes-Mallows Index (FMI). Details can be found in Appendix G.

5.4 Experimental Setup

For the text modality, we utilize the pre-trained BERT model from the Huggingface Transformers library (Wolf et al., 2020) and optimize it using the AdamW (Loshchilov and Hutter, 2019) optimizer. It is important to note that all the baselines utilize the same backbone for each of the three modalities for a fair comparison. In our experiments, the multimodal data employed for pre-training and training adhere to a consistent distribution and characteristics, and no external data is used for pre-training.

We configure the sequence lengths $L_{\text{T}}, L_{\text{V}}, L_{\text{A}}$ for MIntRec, MELD-DA, and IEMOCAP-DA datasets to (30, 230, 480), (70, 250, 520), and (44, 230, 380), respectively. The threshold t is

	Methods	NMI	ARI	ACC	FMI	Avg.
MIntRec	w/o Step 1	31.73	7.70	23.96	13.27	19.17
	Random (Step 2)	45.51	21.32	40.18	26.28	33.32
	SCL (Step 3)	40.44	15.91	32.36	21.63	27.59
	Step 1&K-Means++	42.16	16.31	35.46	21.39	28.83
	Step 1&UCL	47.33	22.72	43.55	27.53	35.28
	Step 1&MSE	16.47	0.69	15.78	8.50	19.55
	UMC	**49.26**	**24.67**	43.73	**29.39**	**36.76**
M-DA	w/o Step 1	10.68	5.89	23.31	20.99	15.22
	Random (Step 2)	21.05	19.05	35.00	33.01	27.03
	SCL (Step 3)	18.11	10.77	28.02	23.94	20.21
	Step 1&K-Means++	19.45	10.29	25.43	21.88	19.26
	Step 1&UCL	21.08	18.77	33.25	31.59	26.17
	Step 1&MSE	5.26	0.89	24.65	26.39	14.30
	UMC	**23.22**	**20.59**	**35.31**	**33.88**	**28.25**
I-DA	w/o Step 1	9.85	4.19	25.26	20.42	14.93
	Random (Step 2)	23.39	19.20	32.59	31.47	26.66
	SCL (Step 3)	16.50	10.08	27.07	23.80	19.36
	Step 1&K-Means++	14.32	6.06	21.80	18.16	15.09
	Step 1&UCL	21.69	13.24	26.05	25.18	21.54
	Step 1&MSE	5.46	-2.29	24.12	25.33	13.15
	UMC	**24.16**	**20.31**	**33.87**	**32.49**	**27.71**

Table 3: Ablation studies on the three datasets.

incremented by Δ of 0.05. For the selection of optimal K_{near}, we configure $L = 0.1$, $\Delta' = 0.02$, and $u = 10$. The learning rates are 2e-5 and (3e-4, 2e-4, 5e-4) for pre-training and training stages of MIntRec, MELD-DA, and IEMOCAP-DA datasets. The temperature parameters τ_1, τ_2, and τ_3 are set at 0.2, (1.4, 20, 6) and (1, 20, 6) for these datasets, respectively. A detailed hyper-parameter sensitivity analysis is provided in Appendix H. We use a training batch size of 128 and report an average performance over five random seeds of 0-4.

5.5 Results

Table 2 shows the results on the multimodal semantics discovery task. UMC (Text), a variant of our method utilizing only the text modality, exhibits comparable or superior performance to existing state-of-the-art methods across most clustering metrics. This indicates our proposed clustering method is highly effective with innovative high-quality sampling and two-step representation learning strategies.

Compared to UMC (Text), UMC incorporates non-verbal modalities and demonstrates significant and consistent improvements of 1-3%, 2-4%, and 1-4% on the MIntRec, MELD-DA, and IEMOCAP-DA datasets, respectively, across all clustering metrics. These results highlight the importance of non-verbal modalities and illustrate that our data augmentation strategy can effectively model multimodal interactions, thereby enhancing the learning of robust representations conducive to clustering. Remarkably, UMC achieves notable increases of 2-7%, 2-7%, and 2-8% in ARI, ACC, and FMI
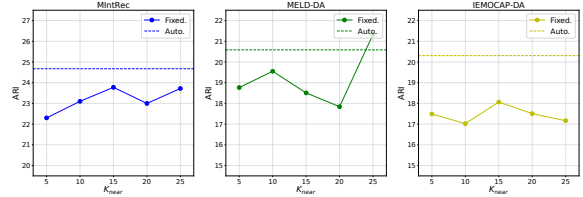
Figure 4: Automatic vs. fixed K_{near} selection strategy.

across all datasets, verifying the ability of UMC to capture complex multimodal semantics in our challenging task. A case study for error analysis is detailed in Appendix I.

6 Discussion

6.1 Ablation Studies

We conduct extensive ablation studies and show the results in Table 3. (1) w/o Step 1: Removing Step 1 results in performance drops of 11-14%, 12-15%, and 8-15% across the MIntRec, MELD-DA, and IEMOCAP-DA datasets, emphasizing the importance of our proposed non-verbal modality masking strategy in enhancing subsequent clustering.

(2) Random (Step 2): To assess the impact of our high-quality sampling strategy in Step 2, we replace it with random sampling (i.e., randomly selecting the top-*t* percent of samples from each cluster). This change leads to average score decreases of 3.42%, 1.22%, and 1.05% on the clustering metrics, highlighting that carefully selected high-quality samples are pivotal in guiding the learning of multimodal representations.

(3) SCL (Step 3): To evaluate the two-step learning approach in Step 3, we remove the unsupervised contrastive learning loss (UCL), resulting in more significant decreases of 6-11% across all three datasets.

(4) Step 1 & other strategies (K-Means++, UCL, MSE): Since the high-quality sampling strategy (Step 2) works in conjunction with multimodal representation learning (Step 3), we experiment with alternative strategies and observe their performance. Initially, applying K-Means++ directly after Step 1 leads to dramatic drops of over 10% across all three datasets. Then, implementing UCL after Step 1 still brings noticeable decreases of 1.48%, 1.08%, and 6.17% in average clustering metric scores. Lastly, we apply a mean squared error (MSE) loss between each sample feature and its corresponding cluster centroid, resulting in extremely low performance with decreases of over 20% in ARI scores. These
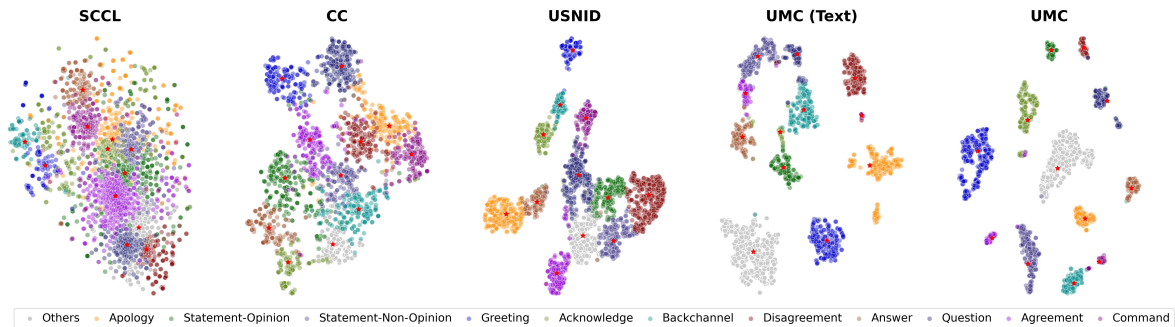
Figure 5: Visualization of representations on the IEMOCAP-DA dataset.

ablation studies further validate the effectiveness of each component in our proposed UMC algorithm.

6.2 Effect of the K_{near} Selection Strategy

In Section 4.3.2, we introduce an automatic method for determining the optimal K_{near} for each cluster. To demonstrate its efficacy, we compare it with a fixed K_{near} approach, where the fixed value varies from 5 to 25 in increments of 5. We then evaluate the performance using ARI scores.

As shown in Figure 4, the automatic K_{near} selection strategy outperforms all fixed K_{near} settings, except for K_{near}=25 in the MELD-DA dataset, which only shows a slight decrease. However, this particular hyper-parameter shows a substantial decrease in the other two datasets. The reason is that the fixed strategy struggles with the imbalanced data distribution across clusters of varying sizes, whereas our approach adapts K_{near} to the unique characteristics of each cluster. Importantly, this approach obviates the need for extensive manual hyper-parameter tuning while still ensuring excellent performance.

6.3 Visualization

Figure 5 uses t-SNE (Maaten and Hinton, 2008) to visualize representations on the IEMOCAP-DA dataset, with additional results provided in Appendix J. SCCL exhibits substantial overlap among intent classes. CC displays more compact clusters, yet still presents implicit cluster boundaries. USNID shows clear cluster boundaries, but the different clusters are close in the feature space and difficult to discern. UMC (Text) demonstrates the most distinct cluster boundaries among text-based baselines, highlighting the robustness of the representations learned through our clustering algorithm. When incorporating non-verbal modalities, the multimodal representations learned by UMC

reveal that each cluster is both compact and well-separated from others, verifying its efficacy.

7 Conclusions

This paper introduces the multimodal semantics discovery task and proposes a novel unsupervised multimodal clustering (UMC) method to address this critical challenge. UMC effectively utilizes non-verbal modalities for semantics discovery by constructing positive multimodal data augmentations. Besides, it proposes a novel high-quality sample selection mechanism and a two-step representation learning strategy.

We conduct extensive experiments on both multimodal intent and dialogue act benchmark datasets. UMC achieves remarkable improvements of 2-6% in standard clustering metrics compared to state-of-the-art clustering algorithms. Further analyses demonstrate the effectiveness of each component and the robustness of the learned representations conducive to clustering. We believe this work makes significant progress in this area and provide a solid foundation for related research.

8 Limitations

There are two limitations in this work. Firstly, given the complexity of real-world multimodal intent datasets, the achieved clustering performance still suggests significant potential for further improvements. Secondly, while this study establishes a foundational approach for automatically determining the K_{near} parameter, there is scope for exploring diverse methodologies within this automatic selection mechanism.

9 Acknowledgements

References

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. In *Proc. of NeurIPS*, pages 9758–9770.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: the advantages of careful seeding. In *Proc. of SODA*, pages 1027–1035.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. pages 12449–12460.

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.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proc. of ECCV*, pages 132–149.

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

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proc. of ICML*, pages 1597–1607.

Jackie Chi Kit Cheung and Xiao Li. 2012. Sequence clustering and labeling for unsupervised query intent discovery. In *Proc. of WSDM*, pages 383–392.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.

Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, and Abdelaziz Bouras. 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, pages 267–279.

K Chidananda Gowda and G Krishna. 1978. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, 10(2):105–112.

Dilek Hakkani-Tür, Yun-Cheng Ju, Geoffrey Zweig, and Gokhan Tur. 2015. Clustering novel intents in a conversational interaction system with semantic parsing. In *Proc. of Interspeech*.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proc. of EMNLP*, pages 9180–9192.

Iryna Haponchyk, Antonio Uva, Seunghak Yu, Olga Uryupina, and Alessandro Moschitti. 2018. Supervised clustering of questions into intents for dialog system applications. In *Proc. of EMNLP*, pages 2310–2321.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proc. of ACM MM*, pages 1122–1131.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of CVPR*, pages 770–778.

Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *Proc. of CVPR*, pages 9248–9257.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proc. of EMNLP*, pages 7837–7851.

Shudong Huang, Ivor W Tsang, Zenglin Xu, Jiancheng Lv, and Quanhui Liu. 2021. Cdd: Multi-view subspace clustering via cross-view diversity detection. In *Proc. of ACM MM*, pages 2308–2316.

Rajat Kumar, Mayur Patidar, Vaibhav Varshney, Lovekesh Vig, and Gautam Shroff. 2022. Intent detection and discovery from user logs via deep semi-supervised contrastive clustering. In *Proc. of NAACL*, pages 1836–1853.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive clustering. In *Proc. of AAAI*, pages 8547–8555.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proc. of AAAI*, pages 8360–8367.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of ICCV*, pages 10012–10022.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proc. of ICLR*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11):2579–2605.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297.

Adyasha Maharana, Quan Hung Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. 2022. Multimodal intent discovery from livestream videos. In *Proc. of NAACL Findings*, pages 476–489.

Yutao Mou, Keqing He, Yanan Wu, Zhiyuan Zeng, Hong Xu, Huixing Jiang, Wei Wu, and Weiran Xu. 2022. Disentangled knowledge transfer for OOD intent discovery with unified contrastive learning. In *Proc. of ACL*, pages 46–53.

Yutao Mou, Xiaoshuai Song, Keqing He, Chen Zeng, Pei Wang, Jingang Wang, Yunsen Xian, and Weiran Xu. 2023. Decoupling pseudo label disambiguation and representation learning for generalized intent discovery. In *Proc. of ACL*, pages 9661–9675.

Srinivas Bangalore Padmasundari and Srinivas Bangalore. 2018. Intent discovery through unsupervised semantic text clustering. In *Proc. of Interspeech*.

Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. 2019. Comic: Multi-view clustering without parameter selection. In *Proc. of ICML*, pages 5092–5101.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. of ACL*, pages 527–536.

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 *Proc. of ACL*, pages 2359–2369.

Tulika Saha, Dhawal Gupta, Sriparna Saha, and Pushpak Bhattacharyya. 2021a. Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. *Cognitive Computation*, 13:277–289.

Tulika Saha, Aditya Prakash Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. In *Proc. of ACL*, pages 4361–4372.

Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. 2021b. Towards sentiment and emotion aided multi-modal speech act classification in Twitter. In *Proc. of NAACL*, pages 5727–5737.

Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. 2017. A review of clustering techniques and developments. *Neurocomputing*, pages 664–681.

Tao Shi and Shao-Lun Huang. 2023. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proc. of ACL*, pages 14752–14766.

Wenkai Shi, Wenbin An, Feng Tian, Qinghua Zheng, QianYing Wang, and Ping Chen. 2023. A diffusion weighted graph framework for new intent discovery. In *Proc. of EMNLP*, pages 8033–8042.

Mengjing Sun, Pei Zhang, Siwei Wang, Sihang Zhou, Wenxuan Tu, Xinwang Liu, En Zhu, and Changjian Wang. 2021. Scalable multi-view subspace clustering with unified anchors. In *Proc. of ACM MM*, pages 3528–3536.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proc. of AAAI*, pages 8992–8999.

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 *Proc. of ACL*, pages 6558–6569.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*, pages 6000–6010.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12).

Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. 2023. Tackling modality heterogeneity with multiview calibration network for multimodal sentiment detection. In *Proc. of ACL*, pages 5240–5252.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural

language processing. In *Proc. of EMNLP*, pages 38–45.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proc. of ICML*, pages 478–487.

Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proc. of ICML*, pages 3861–3870.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In *Proc. of ACL*, pages 7617–7630.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proc. of ACL*, pages 3718–3727.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proc. of AAAI*, pages 10790–10797.

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 *Proc. of ACL*, pages 2236–2246.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen R McKeown, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang. 2021a. Supporting clustering with contrastive learning. In *Proc. of NAACL*, pages 5419–5430.

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. 2021b. TEXTOIR: An integrated and visualized platform for text open intent recognition. In *Proc. of ACL*, pages 167–174.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. 2024. MIntrec 2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *Proc. of ICLR*.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021c. Discovering new intents with deep aligned clustering. In *Proc. of AAAI*, pages 14365–14373.

Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. 2023. A clustering framework for unsupervised and semi-supervised new intent discovery. *IEEE Transactions on Knowledge and Data Engineering*.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022a. Mintrec: A new dataset for multimodal intent recognition. In *Proc. of ACM MM*, pages 1688–1697.

Yi Zhang, Xinwang Liu, Siwei Wang, Jiyuan Liu, Sisi Dai, and En Zhu. 2021d. One-stage incomplete multi-view clustering via late fusion. In *Proc. of ACM MM*, pages 2717–2725.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. 2022b. New intent discovery with pre-training and contrastive learning. In *Proc. of ACL*, pages 256–269.

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. 2024. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proc. of AAAI*, pages 17114–17122.

Yunhua Zhou, Guofeng Quan, and Xipeng Qiu. 2023. A probabilistic framework for discovering new intents. In *Proc. of ACL*, pages 3771–3784.

A   Limitations of Text-only Information in Multimodal Semantics Discovery

Multimodal information is crucial for semantics discovery, as it encompasses a broader range of communicative cues beyond mere text, such as tone of voice, facial expressions, and body language. These cues significantly enhance human communication by conveying subtle nuances and emotions that text alone cannot fully capture.

To illustrate this, we analyze examples from the MIntRec dataset, particularly focusing on two clusters with intents of *Joke* (top) and *Prevent* (bottom), as detailed in Table 4. In the *Joke* cluster, the first two examples contain text utterances with exaggerated rhetoric, clearly conveying the intent of *Joke*. However, the latter four examples, with the semantics of *Statement-opinion*, *Question*, and *Statement-non-opinion*, are less straightforward. Their real intention become clearer when considering non-verbal cues like exaggerated body language and expressions in a relaxed and happy tone.

Similarly, in the *Prevent* intent cluster, the first two examples with clear negative directives are easily distinguished from text. However, the following three examples, which misleadingly suggest intentions of *Agree*, *Oppose*, and *Inform* when relying solely on text. Here, non-verbal cues like *nodding* and *arm blocking* from body language, combined with a resolute voice of tone, are vital for discovering the real *Prevent* intention.

Hence, incorporating non-verbal modalities is essential in real-world contexts for a comprehensive

understanding of the complex semantics in human language. This shows the importance of leveraging non-verbal modalities when performing semantics discovery.

B Applications of Multimodal Semantics Discovery

Video Content Recommendation: Online short video platforms, such as TikTok, have become globally popular, featuring content that includes text, video, and audio elements provided by content creators. Given the vast volume of videos on the internet, accurately tagging each video to match individual user preferences can be prohibitively expensive. Therefore, discovering potential user intentions from unsupervised multimodal data is crucial. An effective multimodal clustering method can discover user needs and group similar content, significantly improving the relevance of recommendations for content retrieval and search.

Efficient Multimodal Data Annotation: A well-trained multimodal clustering model is invaluable for processing real-world multimodal data. It can quickly create clusters based on similar multimodal characteristics, facilitating the identification and analysis of new patterns. Moreover, it enables the efficient generation of semantic annotations at the cluster level, speeding up the annotation process and reducing the workload compared to instance-level annotation.

Virtual Human: Virtual humans hold significant commercial value for many businesses, with some companies promoting custom-designed robots as flagship products. However, effective virtual humans must be able to accurately capture human intentions from various signals, including natural language, body language, facial expressions, and vocal tone. Given that data from real-world human-machine interactions are often unsupervised, it is vital for virtual humans to discern potential user needs from clustered data. This capability allows them to offer better performance and interact with humans in a more natural and fluent manner.

Overall, multimodal semantics discovery opens up new possibilities for the analysis and interpretation of unsupervised multimodal data, which is increasingly prevalent in our digital communication era.

C Additional Related Works

C.1 Multi-view Clustering

Multi-view clustering primarily employs matrix optimization algorithms such as CDD (Huang et al., 2021), COMIC (Peng et al., 2019), OS-LF-IMVC (Zhang et al., 2021d), and SMVSC (Sun et al., 2021). These algorithms utilize graphical or spatial methods to mathematically divide clustering into several sub-tasks and then iteratively optimize the subtask matrices. However, multi-view clustering may become inefficient when processing high-dimensional data, and its time cost can increase at an ultra-linear rate with larger datasets. Besides, the design of optimization objectives in multi-view clustering methods presents certain challenges and does not always guarantee favorable results.

In contrast, multimodal clustering, which tends to focus on deep neural network methodologies, can alleviate these difficulties. For example, XDC (Alwassel et al., 2020) clusters two separate modalities and employs cross-modal pseudo-labels as a supervisory signal for model training, effectively utilizing the semantic connections and distinctions between different modalities. DMC (Hu et al., 2019) uses an exponential function approximation to enable differentiable minimum optimization for clustering, drawing data points closer to their cluster center. It is important to note that these methods are limited to bimodal learning rather than accommodating multiple modalities.

C.2 Multimodal Language Analysis

Multimodal language analysis has introduced numerous datasets (Zadeh et al., 2016, 2018; Yu et al., 2020) and multimodal fusion methods (Tsai et al., 2019; Sun et al., 2020; Rahman et al., 2020; Hazarika et al., 2020; Yu et al., 2021; Han et al., 2021; Maharana et al., 2022; Hu et al., 2022; Wei et al., 2023; Yang et al., 2023; Shi and Huang, 2023).

While most research has focused on meta properties like emotions or sentiment, less attention has been given to the content semantics of multimodal utterances. Zhou et al. (2024) has specifically designed a method for multimodal intent recognition, leveraging the text modality to guide the learning of prompts from non-verbal modalities. However, this method is inapplicable for unsupervised scenarios.

To address this, EMOTyDA (Saha et al., 2020) offers dialogue act labels that complement two multimodal emotion datasets (Busso et al., 2008; Poria et al., 2019), and recent studies have ventured
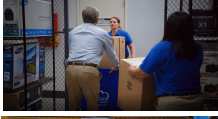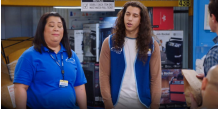
Text	Video	Audio	Require Non-verbal Modalities	Useful Signals
he's, like, a major fox.			✘	Natural Language
you're like one of those monks in tibet.			✘	Natural Language
and you're on the phone.			✔	Tone of Voice, Expressions
and you got that from pants?			✔	Tone of Voice, Expressions
running hard, water bad.			✔	Body Language, Expressions
i can do impressions too.			✔	Tone of Voice, Expressions
okay, bo, stop. all right?			✘	Natural Language
oh, god. sandra, stop, please.			✘	Natural Language
oh, yeah, yeah, yeah, yeah, yeah, yeah, yeah.			✔	Body Language
oh, absolutely not.			✔	Body Language, Natural Language
mom, come on.			✔	Expressions, Tone of Voice

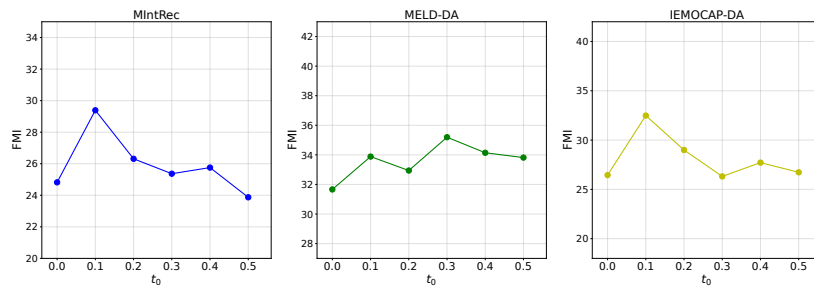Table 4: Real-world examples of *Joke* (top) and *Prevent* (bottom) intent clusters.

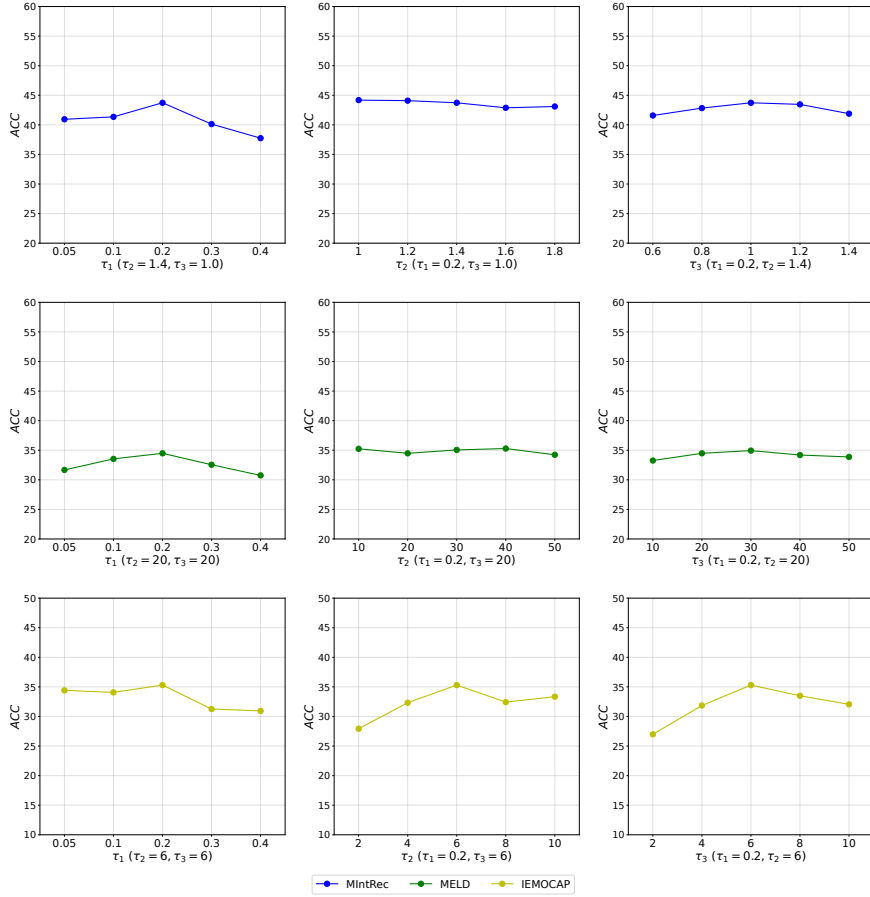Figure 6: Results of clustering with varying initial thresholds t_0.

Figure 7: Sensitivity analysis of τ_1, τ_2, and τ_3 on the three datasets.

Backbones	NMI	ARI	ACC	FMI
ResNet-50 + wav2vec 2.0	47.04	22.54	42.11	27.37
Swin Transformer + WavLM	**49.26**	**24.67**	**43.73**	**29.39**

Table 5: Effect of the multimodal features on the MIntRec dataset.

into multimodal dialogue act classification (Saha et al., 2021b,a). Maharana et al. (2022) introduce a dataset for recognizing operational intents in instructive videos, employing a multimodal cascaded cross-attention late fusion model. MIntRec (Zhang et al., 2022a) provides the first multimodal dataset for conversational intent recognition, using multimodal fusion methods as benchmarks. However, these works depend on supervised learning with provided labels, with few studies in the area of unsupervised multimodal language analysis. Very recently, Zhang et al. (2024) introduces the first large-scale multimodal dataset for both intent recognition and out-of-scope detection in conversations, highlighting the challenges of existing machine learning methods in understanding complex semantics within multimodal utterances.
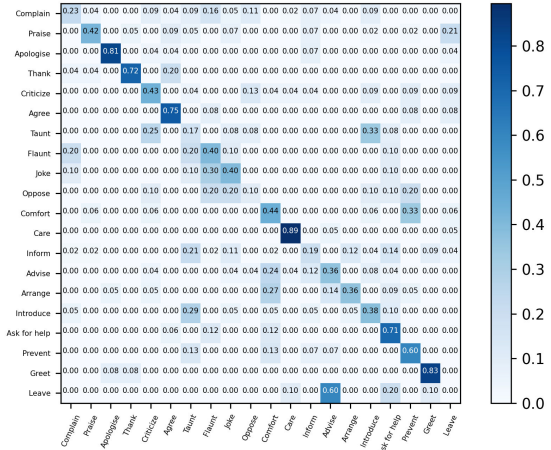

Figure 8: Confusion matrix on the MIntRec dataset.

D Effect of Multimodal Features

In this study, we select the Swin Transformer and WavLM, two state-of-the-art models in computer vision and speech signal processing, as backbones for extracting multimodal features. These models demonstrate superior performance compared to the original features used in the MIntRec (Zhang

et al., 2022a) and EMOTyDA (Saha et al., 2020) datasets. Due to the unavailability of features from EMOTyDA, which poses a challenge for reproduction, we employed features from the MIntRec dataset. MIntRec utilizes ResNet (He et al., 2016) and wav2vec 2.0 (Baevski et al., 2020) for video and audio modalities, respectively.

As illustrated in Table 5, our results show that Swin Transformer and WavLM enhance performance by approximately 2-3% on the MIntRec dataset. This improvement evidences their effectiveness in modeling multimodal representations and capturing the semantics of non-verbal modalities, which are pivotal for cross-modal interactions.

E   Selection of t_0

To select the appropriate parameter t_0 as mentioned in Eq. 4, we vary t_0 from 0.0 to 0.5 at intervals of 0.1 and present the experimental results of the clustering metric FMI on three multimodal intent datasets.

As shown in Figure 6, the clustering results fluctuate with different values of t_0, with t_0=0.1 generally achieving the best performance. Specifically, this value of t_0 achieve the highest performance on the MIntRec and IEMOCAP-DA datasets, and comparable performance on the MELD-DA dataset. This is reasonable because a larger t_0 tends to include more data initially, which may introduce more low-quality data as anchors, thereby hindering the learning of representations conducive to effective clustering.

F   Dataset Specifications and Split Details

Multimodal Intent Dataset: MIntRec (Zhang et al., 2022a) is the premier dataset for multimodal intent recognition in conversation scenarios, spanning text, audio, and video modalities. It comprises 20 intent classes with 2,224 high-quality annotated samples. The original dataset has a 3:1:1 split for training, validation, and testing. As unsupervised clustering does not require the validation set, we merge it with the training set, resulting in a 4:1 ratio between the training and testing sets.

Multimodal Dialogue Act Datasets: We use two large-scale multimodal dialogue act datasets, MELD-DA and IEMOCAP-DA, which are derived from the MELD (Poria et al., 2019) and IEMO-CAP (Busso et al., 2008) datasets, respectively. The EMOTyDA (Saha et al., 2020) dataset provides dialogue act labels for these datasets, encompass-

ing 12 dialogue act classes. We maintain a 4:1 data split ratio for training and testing, consistent with the split used for MIntRec.

G   Evaluation Metrics

Particularly, ACC is calculated by aligning predictions and ground truth using the Hungarian algorithm, as described in (Zhang et al., 2021a, 2023). For NMI, ACC, FMI, the range of possible values is from 0 to 1, while ARI ranges from -1 to 1. Higher values of all these metrics indicate better clustering performance.

The normalized mutual information (NMI) is defined as:

$$\text{NMI}(\mathbf{y}^{gt}, \mathbf{y}^p) = \frac{MI(\mathbf{y}^{gt}, \mathbf{y}^p)}{\frac{1}{2}(H(\mathbf{y}^{gt}) + H(\mathbf{y}^p))}, \quad (12)$$

where $\mathbf{y}^{gt}$ and $\mathbf{y}^p$ are the ground-truth and predicted labels, respectively. $MI(\mathbf{y}^{gt}, \mathbf{y}^p)$ represents the mutual information between $\mathbf{y}^{gt}$ and $\mathbf{y}^p$, and $H(\cdot)$ is the entropy. The mutual information is normalized by the arithmetic mean of $H(\mathbf{y}^{gt})$ and $H(\mathbf{y}^p)$, and the resulting NMI values fall within the range of [0, 1].

The adjusted Rand index (ARI) is defined as:

$$\text{ARI} =$$
$$\frac{\sum_{i,j} \binom{n_{i,j}}{2} - [\sum_i \binom{u_i}{2} \sum_j \binom{v_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{u_i}{2} + \sum_j \binom{v_j}{2}] - [\sum_i \binom{u_i}{2} \sum_j \binom{v_j}{2}]/\binom{n}{2}}, \quad (13)$$

where $u_i = \sum_j n_{i,j}$, and $v_j = \sum_i n_{i,j}$. n is the number of samples, and $n_{i,j}$ is the number of samples that have both the ith predicted label and the jth ground-truth label. The values of ARI fall within the range of [-1, 1].

The accuracy (ACC) is defined as:

$$\text{ACC}(\mathbf{y}^{gt}, \mathbf{y}^p) = \max_m \frac{\sum_{i=1}^n \mathbb{I}\left\{y_i^{gt} = m(y_i^p)\right\}}{n}, \quad (14)$$

where m is a one-to-one mapping between the ground-truth label $\mathbf{y}^{gt}$ and predicted label $\mathbf{y}^p$ of the ith sample. The Hungarian algorithm efficiently obtains the best mapping m. The values of ACC range from [0, 1].

FMI (Fowlkes-Mallows Index) is defined as:

$$\text{FMI}(\mathbf{y}^{gt}, \mathbf{y}^p) = \frac{\text{TP}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})}}, \quad (15)$$
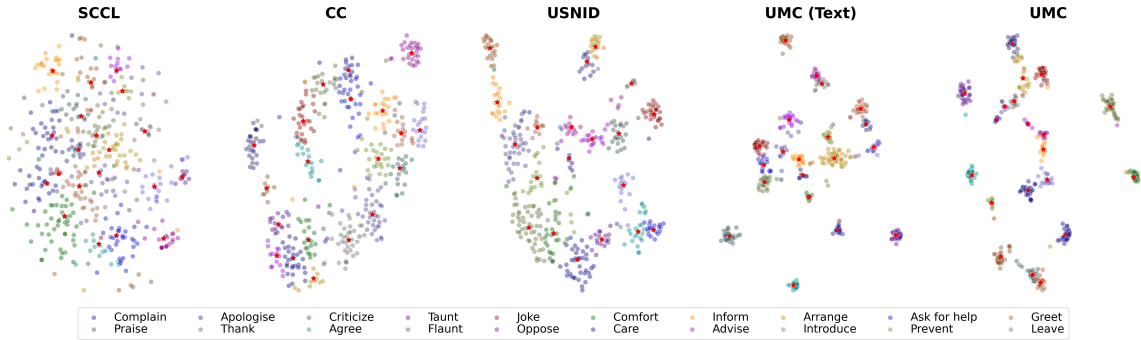
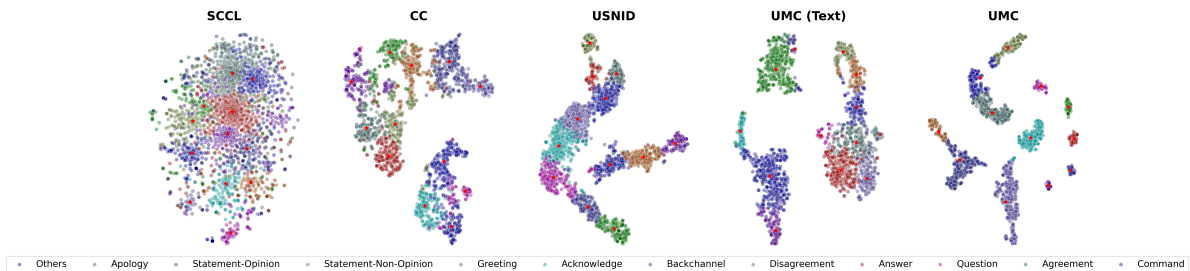Figure 9: Visualization of representations on the MIntRec dataset.

Figure 10: Visualization of representations on the MELD-DA dataset.

where $\mathbf{y}^{gt}$ and $\mathbf{y}^p$ are the ground-truth and predicted labels, respectively. TP represents the number of true positive instances, FP is the number of false positive instances, and FN is the number of false negative instances. The FMI is calculated as the ratio of true positive instances to the geometric mean of the product of false positives and false negatives. The values of FMI range from 0 to 1, with higher values indicating better clustering performance.

H Hyper-parameter Sensitivity Analysis

We conduct a sensitivity analysis on the key hyper-parameters τ_1, τ_2, and τ_3, essential for multimodal unsupervised pre-training and representation learning. The results are displayed in Figure 7.

Initially, with τ_2 and τ_3 set to optimal values, we explore the impact of varying τ_1 on multimodal clustering. The optimal value for τ_1 is 0.2 across all three datasets, and any deviation from this value results in a performance decline. Next, keeping τ_1 and τ_3 constant, we find that the optimal settings for τ_2 are 1.4, 20, and 6 for the MIntRec, MELD-DA, and IEMOCAP-DA datasets, respectively, with a slight performance fluctuating on the MELD-DA dataset. Finally, by fixing τ_1 and τ_2, we observe that τ_3 tends to cause more significant performance

fluctuations on the IEMOCAP-DA dataset. Following a pattern similar to the other hyper-parameters, τ_3 shows local optima at values of 1, 20, and 6 for the respective datasets.

I Error Analysis

Utilizing Hungarian alignment between predictions and ground truth, we present the confusion matrix in Figure 8. Initially, we note that several classes with simpler semantics, such as *Apologise*, *Thank*, *Care*, and *Greet*, achieve near or over 90% accuracy. This is reasonable, as they can be easily identified using text information, and the addition of non-verbal information maintains this advantage.

However, for moderately difficult intent classes like *Advise*, *Ask for Help*, *Prevent*, and *Agree*, which achieve around 60% accuracy, non-verbal cues such as *nodding* and *gestures* are essential to infer the true intent. On the other hand, some intent classes, including *Oppose*, *Complain*, *Taunt*, *Flaunt*, and *Joke*, are particularly challenging, with very few or even no samples accurately clustered. These classes often exhibit complex semantics that require a well-combined analysis of different modalities to accurately interpret real human intentions, resulting in poor performance in multimodal clustering. The overall modest performance also

indicates significant room for improvement in the field of unsupervised multimodal clustering.

J   Representation Visualization

Besides the visualized representations on the IEMOCAP-DA dataset, as introduced in Figure 5, we visualize the representations on the MIntRec and MELD-DA datasets. These are respectively illustrated in Figure 9 and Figure 10. SCCL struggles to effectively learn cluster-level features, resulting in a trivial case of discrete points. CC performs better, forming cluster shapes that are relatively easy to distinguish. USNID shows more compact cluster boundaries, but the boundaries between adjacent clusters are still somewhat indistinct. UMC (Text) performs the best among the text baselines but still has difficult clusters that are close to others. In contrast, our proposed UMC method displays explicit decision boundaries between different clusters, with intra-cluster cohesion being more compact compared to the other methods.