Unsupervised Semantic Discovery via Global and Local Semantic Alignment in Multimodal Clustering

Zhengzhong Zhu[1], Pei Zhou[1], Weihong Du[1], Shiquan Min[1], Jiangping Zhu[1]*

[1]College of Computer Science, Sichuan University, Chengdu, China
{zjp16,zhoupei}@scu.edu.cn, {zhuzhengzhong, duweihong,minshiquan}@stu.scu.edu.cn

Abstract

Unsupervised multimodal semantic discovery aims to learn discriminative representations from multimodal data. However, existing methods suffer from two key limitations. First, they only align instances across modalities without modeling semantic-level consistency, which fails to mitigate semantic bias caused by the gaps among feature distributions of multiple modalities. Second, they inevitably generate incorrect negative pairs during contrastive learning, pushing semantically similar samples apart. To address these challenges, we propose GLAD (Global and Local semantic Alignment for unsupervised multimodal semantic Discovery), which aligns multimodal data at both global and local semantic levels. At the global level, GSA integrates multi-modal features into a shared space and employs joint clustering via optimal transport to capture common semantic patterns while mitigating cross-modality semantic bias. At the local level, LSA adaptively weights samples within each cluster based on their semantic importance, alleviating the effect of incorrect negative pairs. Through the joint optimization of GSA and LSA, GLAD effectively captures both the global semantic structure and the local semantic nuances of multimodal data. Extensive experiments on three benchmark datasets demonstrate GLAD significantly outperforms state-of-the-art methods, with an average improvement of 3.22%.

Introduction

Multimodal semantic discovery seeks to automatically uncover the latent semantic structure of human utterances by leveraging the complementary strengths of multiple modalities. As applications like dialogue systems(Min et al. 2021; Vedula et al. 2019), customer queries(Zhang et al. 2022b), and human-computer interaction(Song et al. 2023; Fang et al. 2023) become increasingly multimodal, the demand for robust and generalizable semantic discovery systems has reached an unprecedented level.

Recent advances in multimodal language analysis have produced numerous supervised methods for understanding complex utterances (Saha et al. 2021; Zhang et al. 2022a), while unsupervised (Haponchyk et al. 2018; Zhang et al. 2023) and semi-supervised (Zhang et al. 2022b; Zhou, Quan, and Qiu 2023) approaches have been proposed as more scalable alternatives by framing semantic discovery as a clustering problem. However, these approaches remain largely text-only and fail to fully exploit non-verbal cues such as prosody, gesture, and visual context, leaving the challenge of effective semantic discovery in truly multimodal, unlabeled environments unsolved. As the first unsupervised multimodal clustering algorithm for semantic discovery, UMC(Zhang et al. 2024) leverages non-verbal modalities through a high-quality sample selection strategy and an iterative representation learning mechanism. It first obtains initial representations via multimodal data augmentation and unsupervised pretraining, then dynamically selects high-quality samples for supervised contrastive learning, alternating with unsupervised contrastive learning on low-quality samples. This design enables UMC to produce accurate and compact clustering results in a fully unsupervised manner. However, UMC suffers from two critical limitations that undermine the quality and reliability of semantic discovery: it fails to effectively mitigate semantic bias caused by the gaps among feature distributions of multiple modalities, and it generates incorrect negative pairs during contrastive learning, which forces semantically similar samples apart.

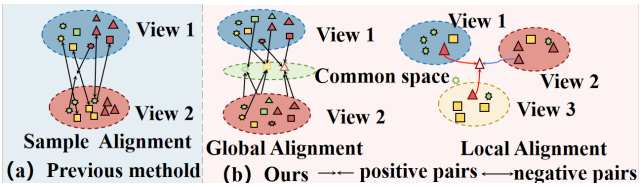


Figure 1: Illustration of the difference between UMC and GLAD. UMC directly aligns cross-modality representations without mitigating the semantic bias caused by the gaps among feature distributions of multiple views and treats all samples equally regardless of their semantic importance. In contrast, GLAD leverages joint clusters in the common space as a global semantic bridge to establish meaningful correlations between modalities, and incorporates a local semantic alignment module to adaptively weight samples based on their semantic significance. This design effectively mitigates semantic bias and achieves semantic-aware cross-modality alignment for robust multi-modal clustering.

*Corresponding author

To address these two issues, we propose GLAD, a novel framework for unsupervised multi-modal semantic discovery. GLAD consists of two complementary modules: Global Semantic Alignment (GSA) and Local Semantic Alignment (LSA). Specifically, GSA aims to mitigate semantic bias by integrating multi-modal features into a unified shared space and performing joint clustering. This allows GSA to capture coherent global semantic patterns shared across all modalities and propagate them back to each view via optimal transport, enabling semantic-aware alignment beyond instance-level consistency. Although UMC implicitly constructs a shared space through feature concatenation for multi-view representation learning, it only explores instance-level consistency between the concatenated representation and specific view representations, as illustrated in Fig.1. However, Although GSA mitigates cross-modal semantic bias at the global level and captures the global semantic patterns shared across modalities, it still faces a challenge—semantic mismatches within specific clusters, especially due to incorrect negative sample selection. Therefore, we introduce LSA, which further refines the alignment at a finer granularity by adaptively weighting samples within each cluster according to their semantic importance. This helps suppress unreliable samples and mitigate the impact of incorrect negative pairs, enhancing the robustness and discriminability of the learned representations. Through the joint optimization of GSA and LSA, GLAD effectively captures both global semantic structure and local semantic nuances in multimodal data, achieving more accurate, coherent, and reliable semantic discovery. Our main contributions are summarized as follows:

1. We analyze the limitations of existing methods for unsupervised multimodal semantic discovery, particularly their inability to effectively bridge the gaps between modalities and learn shared semantics across them.
2. We propose GLAD, a novel framework that introduces two key modules: Global Semantic Alignment (GSA), which aligns global semantic patterns across modalities by jointly clustering in a shared latent space, and Local Semantic Alignment (LSA), which refines alignment by adaptively weighting samples within clusters to reduce semantic bias and incorrect negative pairs.
3. We conduct extensive experiments on benchmark multimodal datasets, demonstrating that GLAD achieves state-of-the-art clustering performance.

Related Work

Unsupervised Multimodal Semantic Discovery

Intent discovery is a key challenge in natural language processing, and many clustering methods have been proposed to tackle this problem. Early approaches (Hakkani-Tür et al. 2015; Haponchyk et al. 2018) relied on weak supervision signals to assist clustering, but struggled to capture high-level semantics in text. More recent methods (Lin, Xu, and Zhang 2020; Zhang et al. 2021b; Mou et al. 2022, 2023; Zhou, Quan, and Qiu 2023; Shi et al. 2023; Peng et al. 2025) leveraged limited labeled data to guide feature learning for clustering, while unsupervised (Cheung and Li 2012; Padmasundari and Bangalore 2018; Haponchyk et al. 2018; Zhang et al. 2023) and semi-supervised (Lin, Xu, and Zhang 2020; Zhang et al. 2021b, 2022b; Zhou, Quan, and Qiu 2023) approaches have been proposed as more scalable alternatives by framing semantic discovery as a clustering problem. However, these methods remain largely text-only and fail to fully exploit non-verbal cues such as prosody, gesture, and visual context, resulting in poor performance in truly multimodal, unlabeled environments. USNID (Zhang et al. 2023) proposed a novel centroid-based mechanism combined with a pre-training strategy, achieving substantial improvements over previous methods. Nevertheless, USNID still performs poorly when handling multimodal data. To address this, MCN (Chen et al. 2021) introduced a unified representation framework and applied cross-modal contrastive loss during clustering. As a pioneering work, UMC (Zhang et al. 2024) applied multimodal clustering to semantic recognition. Specifically, UMC constructs positive and negative sample pairs by combining features from different modalities, and adopts a two-stage training process where high-quality samples are selected using K-means++ to generate pseudo-labels for supervised training, achieving promising results. However, UMC only exploits the consistency of information between sample.

Multimodal Alignment

Multimodal alignment aims to learn semantically consistent representations across different modalities(Zhu et al. 2024; Guan et al. 2025b,c,a). Typical solutions include learning a shared latent space through cross-modal contrastive objectives (Gao et al. 2024), cross-attention mechanisms (Hu et al. 2024; Tsai et al. 2019; Yang et al. 2022), or cross-modal reconstruction (Liu et al. 2024c; Tian et al. 2022; Zeng et al. 2024). However, in unsupervised scenarios, the lack of label information makes alignment more challenging. Current unsupervised approaches commonly adopt a shared representation paradigm, encouraging modal consistency through instance-level contrastive learning. For example, UMC (Zhang et al. 2024) constructs positive and negative pairs by combining features from different modalities, and applies contrastive learning to discover consistent semantic clusters. While these methods achieve cross-modal consistency to some extent, they often overlook modality-specific characteristics, and tend to enforce over-alignment that degrades the distinctiveness of each modality. In contrast, our work explicitly addresses this issue by introducing global and local semantic alignment, which balances cross-modal consistency and intra-modal specificity.

Problem Formulation

A multimodal intent or dialogue act dataset $\mathcal{D}_{\text{mm}} = \{(s_i^{\text{T}}, s_i^{\text{A}}, s_i^{\text{V}}) | y_i \in \mathcal{I},\ i = 1, \ldots, N\}$, where each instance i consists of multimodal utterances, including text s_i^{T}, audio s_i^{A}, and video s_i^{V}. Here, N denotes the total number of instances. The true label y_i, corresponding to one of the intent or dialogue act classes $\mathcal{C} = \{C_i\}_{i=1}^{K_C}$, remains hidden during both training and validation, and is only accessible during testing. The number of classes is denoted by K_C. The ob-

jective is to train a multimodal neural network $\mathcal{F}(\cdot)$ to learn multimodal representations z that are effective for clustering. These representations are then used to partition the set $\{s_i\}_{i=1}^N$ into K_C distinct clusters.

Method

Global Semantic Alignment

We leverage Optimal Transport (OT) to align semantically related instances by bringing their representations closer in the shared space. Given two simplex vectors $\boldsymbol{\alpha} \in \Sigma_r$ and $\boldsymbol{\beta} \in \Sigma_c$ (where $\Sigma_r := \{\mathbf{x} \in \mathbb{R}_+^r \mid \mathbf{x}^\top \mathbf{1}_r = 1\}$), the transportation polytope is defined as:

$$\mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left\{\mathbf{Q} \in \mathbb{R}_+^{r \times c} \mid \mathbf{Q}\mathbf{1}_c = \boldsymbol{\alpha},\ \mathbf{Q}^\top \mathbf{1}_r = \boldsymbol{\beta}\right\}. \quad (1)$$

OT seeks a coupling matrix $\mathbf{Q}$ that maximizes similarity:

$$\mathrm{OT}(\mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \underset{\mathbf{Q} \in \mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\operatorname{argmax}} \operatorname{Tr}\left(\mathbf{Q}^\top \mathbf{S}\right), \quad (2)$$

where $\mathbf{S}$ is the similarity matrix. We adopt the entropic regularized OT formulation:

$$\mathrm{OT}^\epsilon(\mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \underset{\mathbf{Q} \in \mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\operatorname{argmax}} \operatorname{Tr}\left(\mathbf{Q}^\top \mathbf{S}\right) + \epsilon H(\mathbf{Q}), \quad (3)$$

where $H(\mathbf{Q}) = -\sum_{i,j} Q_{ij} \log Q_{ij}$ is the entropy term and $\epsilon > 0$ is the regularization coefficient. Eq. (3) admits a unique solution of the form: $\mathbf{Q}^* = \mathrm{Diag}(\mathbf{u}) \exp\left(\mathbf{S}/\epsilon\right) \mathrm{Diag}(\mathbf{v})$, where $\mathbf{u}$ and $\mathbf{v}$ are computed efficiently via Sinkhorn's algorithm (Cuturi 2013). Finally, our formulation can be summarized as:

$$\mathrm{OT}^\varepsilon(\mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \underset{\mathbf{Q} \in \mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\operatorname{argmax}} \operatorname{Tr}\left(\mathbf{Q}^\top \mathbf{S}\right) + \varepsilon H(\mathbf{Q}), \quad (4)$$

with $\mathbf{Q}^*$ obtained in closed form. To achieve consistent and discriminative semantic representation across multiple modals, we introduce an Optimal Transport-based semantic learning strategy that explicitly models the association between instances and semantic clusters in a shared common space. We begin by constructing a common semantic space through multi-modal feature aggregation. Specifically, for the i-th sample, we compute the averaged embedding:

$$\mathbf{h}_i^c = \frac{1}{M} \sum_{m=1}^{M} \mathbf{h}_i^m, \quad (5)$$

where $\mathbf{h}_i^m$ is the feature representation of the i-th sample in the m-th modal, and $\mathbf{h}_i^c$ denotes its aggregated embedding. This aggregation serves to capture both shared and complementary semantics from all M modals. We then perform k-means clustering over $\{\mathbf{h}_i^c\}_{i=1}^N$ to derive K joint clusters $\{\mathbf{c}_j\}_{j=1}^K$, which represent the dominant semantic patterns in the common space. These clusters are shared across all modals and serve as global semantic anchors. To establish a fine-grained semantic relationship between the samples and the joint clusters, we employ entropic regularized Optimal Transport (OT) to obtain a soft assignment (transport plan) for each modal. For the m-th modal, the OT formulation is:

$$\mathbf{Q}^{m*} = \mathrm{OT}^\varepsilon(\mathbf{S}^m, \tfrac{1}{B}\mathbf{1}_B, \tfrac{1}{K}\mathbf{1}_K), \quad (6)$$

where $\mathbf{S}^m = \mathbf{H}^m \mathbf{C}^\top \in \mathbb{R}^{B \times K}$ denotes the cosine similarity matrix between the mini-batch samples $\mathbf{H}^m$ in the m-th modal and the cluster centers $\mathbf{C}$. Here, B is the batch size and K is the number of clusters. The OT plan $\mathbf{Q}^{m*} \in \mathbb{R}^{B \times K}$ satisfies marginal constraints to ensure balanced assignments. To enhance interpretability and enable probabilistic reasoning, we normalize the OT plan via softmax along the column dimension:

$$\widetilde{\mathbf{Q}}^{m*} = \begin{pmatrix} \widetilde{Q}_{11}^{m*} & \cdots & \widetilde{Q}_{1K}^{m*} \\ \vdots & \ddots & \vdots \\ \widetilde{Q}_{B1}^{m*} & \cdots & \widetilde{Q}_{BK}^{m*} \end{pmatrix}, \quad (7)$$

where each row $\mathbf{q}_i^m$ represents a soft semantic assignment vector for the i-th sample, and each element $\widetilde{Q}_{ij}^{m*}$ indicates the likelihood of the i-th sample being assigned to the j-th cluster. We introduce a KL divergence loss between the assignment vector $\mathbf{q}_i^m$ and the predicted distribution $\mathbf{y}_i^m$:

$$\mathcal{L}_{\mathrm{kl}} = \sum_{m=1}^{M} \sum_{i=1}^{B} \mathrm{KL}(\mathbf{q}_i^m, \mathbf{y}_i^m), \quad (8)$$

which encourages the predicted labels to align with the semantic structure in the common space. In addition, we propose an OT-based semantic alignment loss to explicitly align each sample with its soft cluster assignment. The loss contains both a semantic matching term and an entropy regularization term:

$$\mathcal{L}_{\mathrm{OT}} = \sum_{i=1}^{B} \sum_{j=1}^{K} (1 - S_{ij}^m) \widetilde{Q}_{ij}^{m*} + \sum_{i=1}^{B} \sum_{j=1}^{K} \widetilde{Q}_{ij}^{m*} \ln \widetilde{Q}_{ij}^{m*}, \quad (9)$$

where S_{ij}^m is the cosine similarity between sample $\mathbf{h}_i^m$ and cluster center $\mathbf{c}_j$. The first term promotes semantic closeness, while the second term ensures smooth assignments. Finally, the complete objective for our OT-based semantic learning is defined as:

$$\mathcal{L}_{\mathrm{GSA}} = \mathcal{L}_{\mathrm{kl}} + \mathcal{L}_{\mathrm{OT.}} \quad (10)$$

Local Semantic Alignment

To alleviate the impact of incorrect negative pairs caused by the lack of label information(Zhu et al. 2025b), we introduce the Local semantic alignment(LSA) module, which aligns the fused representation $\mathbf{h}_i^c$ with its corresponding view-specific representation h_i^m while considering their semantic similarity in the cluster space. Specifically, LSA first calculates the similarity matrix of individual modality for all samples as follows:

$$S_{ij}^m = \cos\left(z_i^m, z_j^m\right), \quad (11)$$

where cos is the cosine similarity. Then summing up the similarity matrices of all views and taking the average, we obtain

$$S_{ij} = \frac{1}{M} \sum_{m=1}^{M} S_{ij}^m. \quad (12)$$

The cosine similarity is used to compute the alignment between $\hat{h}_i$ and h_i^m:

$$C\left(\hat{h}_i, h_i^m\right) = \cos\left(\hat{h}_i, h_i^m\right), \quad (13)$$

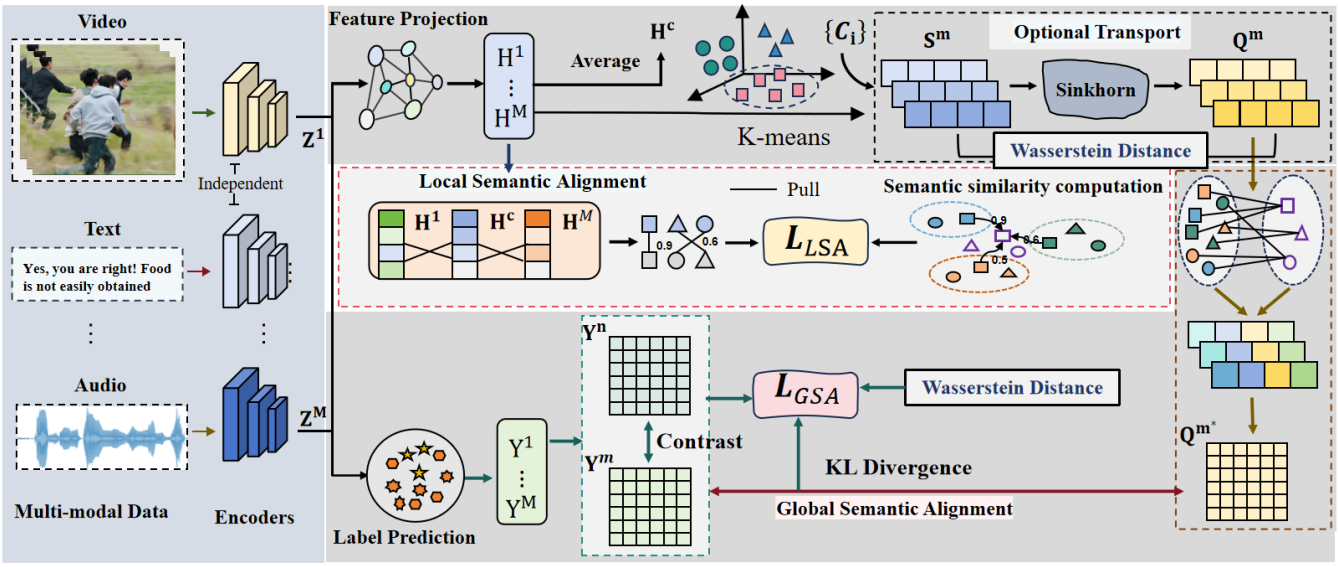


Figure 2: The overall framework of GLAD.

based on the structural similarity s_{ij} computed from the cluster assignments , the LSA loss is defined as:

$$\mathcal{L}_{\text{LSA}} = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{m=1}^{M}\log\frac{e^{C(\hat{h}_i,h_i^m)/\tau}}{\sum_{j=1}^{N} e^{(1-s_{ij})C(\hat{h}_i,h_j^m)/\tau} - e^{1/\tau}}, \tag{14}$$

here, τ is a temperature coefficient controlling the sharpness of distribution. The structural similarity s_{ij} reflects whether samples i and j are from the same semantic cluster. A higher s_{ij} implies stronger semantic weight, thus increasing the contrastive alignment strength between $\hat{h}_i$ and h_j^m. This formulation adaptively reweights the contrastive learning objective to favor alignment among semantically similar samples, effectively capturing fine-grained local semantic structures across modalities(Zhu et al. 2025a).

Joint Training of GLAD

The overall training objective of our GLAD framework includes three loss items:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{GSA}} + \lambda_2\mathcal{L}_{\text{LSA}}, \tag{15}$$

where λ_1 and λ_2 are the weighting factors for semantic learning and weighted contrastive feature learning.

Computational Complexity

We analyze the computational complexity of GLAD by denoting the number of modalities as M and the number of clusters as K. Let B represent the mini-batch size. In each training iteration, computing the global consistency loss $\mathcal{L}_{\text{GSA}}$ requires $\mathcal{O}(MB^2)$ time. Similarly, computing the local consistency loss $\mathcal{L}_{\text{LSA}}$ also requires $\mathcal{O}(MB^2)$ time. For the semantic alignment loss $\mathcal{L}_{\text{LSA}}$, solving the optimal transport problem in Eq. (5) incurs a complexity of $\mathcal{O}(M(BK(I+1)+B+K^2))$, where I is the number of iterations of the Sinkhorn algorithm. Since both B and I are typically much larger than 1 in practice, the overall computational complexity of GLAD during training can be approximated as $\mathcal{O}(M(2B^2+BKI+K^2))$.

Experiment

Datasets and Metrics

Three benchmark datasets for the multimodal semantic discovery task: MIntRec, MELD-DA, and IEMOCAP-DA. Detailed statistics of these datasets are summarized in Appendix B. Following (Fahad et al. 2014; Saxena et al. 2017), we adopt four standard clustering metrics to measure performance: Normalized Mutual Information (NMI), Accuracy (ACC), Adjusted Rand Index (ARI), and Fowlkes-Mallows Index (FMI)(Liu et al. 2025, 2023, 2024a,b).

Experimental Setup

For the text modality, we use the pre-trained BERT model from the Huggingface Transformers library (Wolf et al. 2020), and fine-tune it with the AdamW optimizer (Loshchilov and Hutter 2017). It is important to note that for a fair comparison, all baseline models employ the same backbone architecture for each of the three modalities. In our experiments, the loss weights λ_1 and λ_2 are both set to 0.5, and the temperature coefficient τ is set to 0.05.

Main Results

We compare UMC with the state-of-the-art unsupervised clustering methods from both NLP and CV, as well as multimodal clustering methods. The baselines are as follows:**SCCL** (Zhang et al. 2021a) ,**CC** (Kumar et al. 2022),**USNID** (Zhang et al. 2023),**UMC (Text)(Zhang et al. 2024)**,**MCN** (Chen et al. 2021),**UMC** (Zhang et al. 2024). Table 1 reports the clustering performance of our method compared with several baselines on the MIntRec, MELD-DA (M-DA), and IEMOCAP-DA (I-DA) datasets.

The results show that our method consistently outperforms all baselines across all metrics, achieving an average improvement of about 3.33% over the best-performing baseline UMC on the most challenging I-DA dataset. These results demonstrate that GLAD effectively enhances both global semantic alignment and local consistency, maintaining strong performance and robustness .

	Method	NMI	ARI	ACC	FMI	Avg.
MIntRec	SCCL	45.33	14.6	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	1.7	16.76	10.32	11.76
	UMC (Text)	47.15	25.05	42.46	26.93	34.65
	UMC	49.26	24.67	43.73	29.39	36.76
	Ours	**52.17**	**26.59**	**45.97**	**31.84**	**39.14**
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	20.8	12.16	24.07	23.28	20.08
	MCN	8.34	1.57	18.1	15.31	10.83
	UMC (Text)	19.57	16.29	33.4	30.81	25.02
	UMC	23.22	20.59	35.31	33.88	28.25
	Ours	**25.63**	**24.39**	**37.79**	**36.16**	**30.99**
I-DA	SCCL	21.9	10.9	26.8	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	31.1	25.64
	UMC	24.16	20.31	33.87	32.49	27.71
	Ours	**28.22**	**24.84**	**38.26**	**37.37**	**32.17**

Table 1. Performance comparison across models on multiple datasets. The best performance for each metric is highlighted in bold, and second-best is underlined.

Ablation Studies

Table 2 presents the results of the ablation study, which evaluates the contributions of the GSA and LSA modules on three datasets. As shown, removing either module leads to noticeable performance drops across all metrics, indicating that both modules are essential to the overall effectiveness of the framework. Specifically, removing GSA results in the most significant degradation, particularly on the I-DA dataset, where the average score drops from 32.17 to 18.29, highlighting the critical role of global semantic alignment in modeling complex multimodal semantics. Meanwhile, replacing LSA with w/o s_{ij} (which directly aligns the modality representations of the same sample) results in ACC drops of approximately 2.71%, 2.81%, and 2.13% across datasets, indicating the effectiveness of LSA. On the other hand, removing LSA consistently degrades performance by about 3%–4%, further demonstrating that the local constraint mitigates incorrect negative pairs and improves intra- and inter-cluster structure. Overall, the complete GLAD framework achieves the best performance on all datasets.

Ablation Studies on Optimal Transport in GSA

Table 3 reports the results of the ablation study on the role of optimal transport (OT) within GSA on the MIntRec and M-DA datasets. As shown, replacing OT with a simple cosine similarity (w/ cosine similarity) leads to a noticeable performance drop compared to the full GSA, indicating the importance of OT in modeling accurate sample-to-cluster assignments. Furthermore, completely removing OT (w/o OT) results in even greater degradation across all metrics, highlighting the crucial contribution of global semantic alignment to the overall framework. Specifically, on the MIntRec dataset, the average score decreases from 39.14 (OT) to 35.80 (w/ cosine similarity) and further down to 34.53 (w/o OT). Similar trends are observed on the M-DA dataset, where the average score drops from 30.99 (OT) to 28.70 (w/ cosine similarity) and then to 27.84 (w/o OT). These results demonstrate that OT is essential to the effectiveness of OT by enabling more reliable and semantically meaningful alignments in the common space.

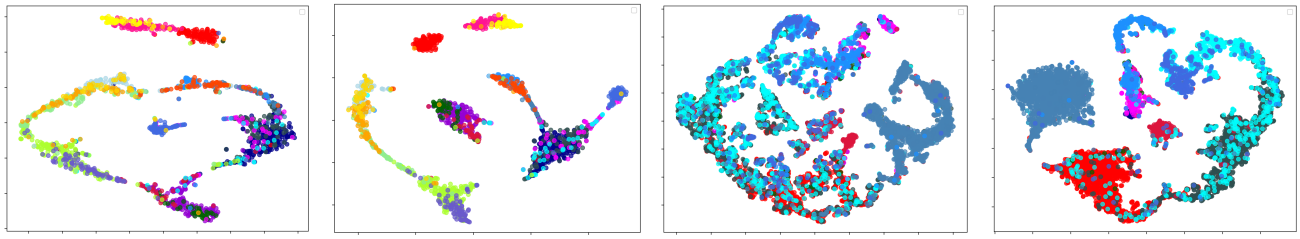
Dataset	Method	NMI	ARI	ACC	FMI	Avg.
MIntRec	w/o GSA	35.73	18.76	26.52	14.57	23.90
	w/o LSA	49.42	23.46	42.51	28.75	36.04
	w/o s_{ij}	50.14	21.37	44.72	29.51	36.44
	GLAD	**52.17**	**26.59**	**45.97**	**31.84**	**39.14**
M-DA	w/o GSA	12.51	11.24	23.37	24.86	17.99
	w/o LSA	23.21	23.75	35.52	34.21	29.17
	w/o s_{ij}	24.32	21.57	36.72	35.14	28.89
	GLAD	**25.63**	**24.39**	**37.79**	**36.16**	**30.99**
I-DA	w/o GSA	15.75	11.59	22.73	23.12	18.29
	w/o LSA	25.54	21.75	31.26	31.79	27.58
	w/o s_{ij}	26.73	22.56	37.44	36.19	30.73
	GLAD	**28.22**	**24.84**	**38.26**	**37.37**	**32.17**

Table 2. Ablation study of GSA and LSA on three datasets.

	Method	NMI	ARI	ACC	FMI	Avg.
MIntRec	w/o OT	49.36	22.74	38.49	27.52	34.53
	w/ cosine similarity	50.21	23.25	40.61	29.14	35.80
	W/ OT	**52.17**	**26.59**	**45.97**	**31.84**	**39.14**
M-DA	w/o OT	21.64	21.81	33.40	34.52	27.84
	w/o cosine similarity	22.17	23.75	34.27	34.62	28.70
	W/ OT	**25.63**	**24.39**	**37.79**	**36.16**	**30.99**

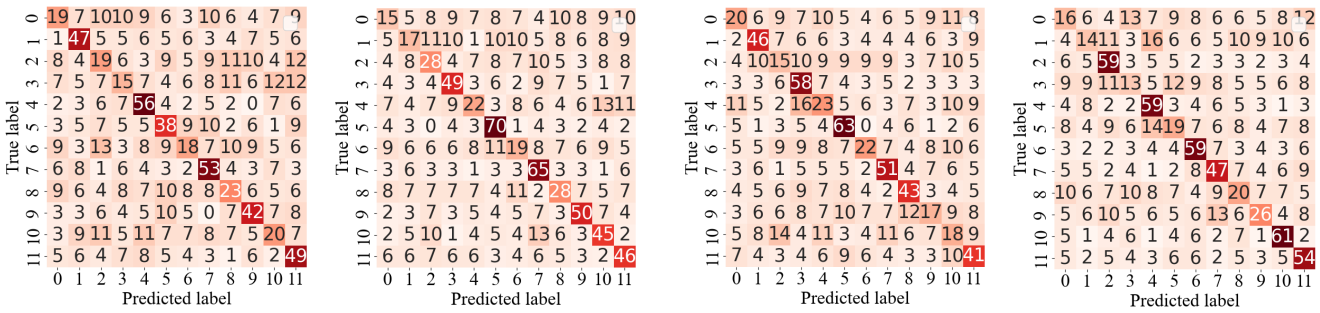
Table 3. Ablation studies of optimal transport on MIntRec and M-DA datasets.

Visualization Figure 3 visualizes the learned representations on the IEMOCAP-DA and I-DA datasets using t-SNE (Van der Maaten and Hinton 2008). Compared to the multimodal representations learned by the UMC model, which exhibit noticeable overlaps between intent categories, our method produces more compact and well-separated clusters in the representation space. This result demonstrates not only the superior ability of our method to capture multimodal semantic distinctions and maintain category separability, but also highlights the advantages of the GLAD framework: the Global Semantic Alignment (GSA) mitigates cross-modal semantic bias, while the Local Semantic Alignment (LSA) suppresses incorrect negative pairs. These two complementary mechanisms jointly enhance the clarity and discriminability of semantic structures in the representation space, leading to significantly improved clustering quality and interpretability. To further demonstrate the effectiveness of GLAD, we compare the confusion matrices generated by GLAD and UMC in Fig. 4. As shown, the diagonal elements in the GLAD matrix are consistently higher and



(a) UMC(MintRec) (b) Ours(MintRec) (c) UMC(IEMOCAP- DA) (d) Ours(IEMOCAP- DA)

Figure 3: Visualization of Learned Representation.



(a) UMC(IEMOCAP-DA) (b) GLAD(IEMOCAP-DA) (c) UMC(MELD-DA) (d) GLAD(MELD-DA)

Figure 4: Confusion matrices generated by the proposed GLAD method and the UMC method.

more concentrated compared to those in the UMC matrix. This improvement stems from the joint effect of GLAD's global semantic alignment and local semantic consistency: the GSA module effectively aligns modality-specific representations to shared semantic clusters, reducing cross-modal semantic bias, while the Local Semantic Alignment module adaptively emphasizes semantically reliable samples within each cluster, suppressing noise from less informative or inconsistent examples.

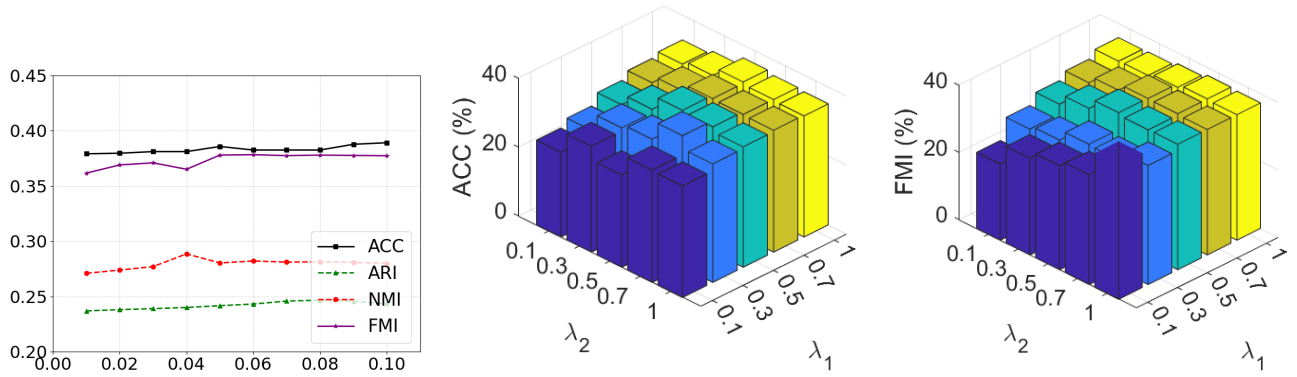
The Almost Stochastic Order (ASO) Test Analysis To further validate the superiority of GLAD, we conduct a statistical significance test using the Almost Stochastic Order (ASO) method (Dror, Shlomov, and Reichart 2019; Ulmer, Hardmeier, and Frellsen 2022). ASO is a statistical approach specifically designed for evaluating deep neural network models and has been widely used for algorithm comparisons in various tasks. In the ASO framework, $\varepsilon_{\text{min}} = 0$ indicates that the evaluated method statistically dominates the compared method, while $\varepsilon_{\text{min}} < 0.5$ suggests almost stochastic dominance. In our experiments, we performed 10 independent runs with different random seeds, comparing GLAD against several recent contrastive multimodal clustering methods. The statistical test results, summarized in Table 4 under a confidence level of $\alpha = 0.05$, demonstrate that GLAD consistently and significantly outperforms all baseline methods with stable performance. The superior performance of GLAD is mainly attributed to two factors. On the one hand, the optimal transport-based semantic learning effectively aligns multi-view semantics in the shared space, extracting complementary information and suppressing redundant information. On the other hand, the local semantic alignment adaptively emphasizes semantically reliable samples and suppresses unreliable ones, mitigating the impact of incorrect negative pairs and improving the discriminability and robustness of the representations.

Evaluation metrics	**MIntRec**		**M-DA**	
	$\varepsilon_{\text{min}}^{ACC}$	$\varepsilon_{\text{min}}^{NMI}$	$\varepsilon_{\text{min}}^{ACC}$	$\varepsilon_{\text{min}}^{NMI}$
GLAD → UMC	6.25e-5	5.33e-5	4.21e-3	6.47e-3
GLAD → MCN	0	0	0	0
GLAD → USNID	0	0	0	0

Table 4. ASO significance test results ($\varepsilon_{\text{min}}^{ACC}$ and $\varepsilon_{\text{min}}^{NMI}$) at $\alpha = 0.05$ confidence level.

Effect of Different Clustering Schemes

We further explore how different clustering approaches influence the final performance. After the training process is complete, both the learned feature projector and label predictor can be used during inference. Based on this, we propose two alternative strategies for generating clustering results during the testing phase: one directly applies k-means to the projected feature representations, while the other derives cluster assignments by taking the argmax of the predicted label vectors. As shown in Fig. 6, the second strategy, using the predicted label vectors, consistently results in better clustering performance. This improvement can be attributed to the explicit semantic supervision provided by GLAD in the shared space during training.



(a) τ (b) ACC (c) NMI

Figure 5: Hyper-parameter on IEMOCAP-DA.

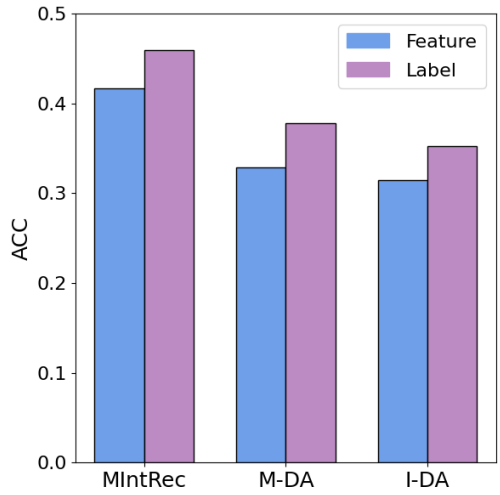


Figure 6: The performance of different clustering schemes.

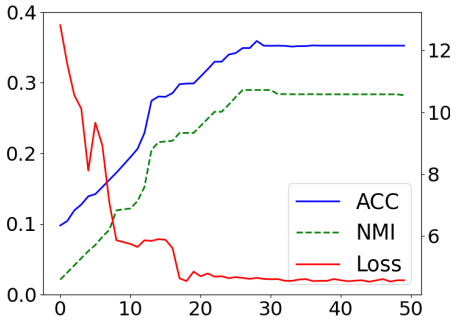


Figure 7: convergence analysis.

Convergence Analysis

We evaluate the convergence behavior of GLAD on the IEMOCAP-DA dataset by tracking the loss and clustering performance across training epochs. As shown in Fig. 7, the loss steadily decreases and stabilizes within 50 epochs, indicating that the model converges efficiently. This fast convergence ensures that the model can quickly learn meaningful semantic representations without prolonged training, which is particularly beneficial for multimodal data where training costs are high due to the large input size and heterogeneous modalities. Moreover, the stable convergence suggests that the global and local alignment objectives are well-balanced.

Hyper-parameter sensitivity analysis

We conduct a sensitivity analysis on the key hyperparameters τ, λ_1, and λ_2 on the IEMOCAP-DA dataset, Fig. 5c and Fig. 5b indicate that λ_1 and λ_2 have a significant impact on model performance, generally exhibiting a trend of first increasing and then decreasing. When λ_1 is too small, global semantic alignment cannot be sufficiently achieved; whereas an overly small λ_2 weakens the local semantic alignment. Therefore, it is crucial to properly balance the two parameters to ensure effective coordination between global and local semantic alignment. In contrast, the temperature parameter τ in Fig.5a causes minor fluctuations in performance, and an appropriate τ can slightly improve the performance.

Conclusion

This paper proposes a novel framework, namely GLAD, which enhances the performance of unsupervised multimodal semantic discovery by learning common semantics through both global and local alignment. Compared to previous work, the most distinctive feature of our approach is that GLAD maps multimodal data into a unified feature space at both global and local levels, offering a more comprehensive perspective for discovering cross-modal common semantics. Experimental results demonstrate that our method effectively uncovers the underlying semantic patterns of multimodal data. A limitation of GLAD, however, is its reliance on complex alignment procedures, which can increase computational cost and limit its efficiency in large-scale or real-time applications. In future work, we will explore more efficient and simplified algorithms to overcome this limitation.

Acknowledgments

This work was supported by the National natural science foundation of China (No. 62571355)

References

Chen, B.; Rouditchenko, A.; Duarte, K.; Kuehne, H.; Thomas, S.; Boggust, A.; Panda, R.; Kingsbury, B.; Feris, R.; Harwath, D.; et al. 2021. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8012–8021.

Cheung, J. C. K.; and Li, X. 2012. Sequence clustering and labeling for unsupervised query intent discovery. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 383–392.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Dror, R.; Shlomov, S.; and Reichart, R. 2019. Deep dominance-how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2773–2785.

Fahad, A.; Alshatri, N.; Tari, Z.; Alamri, A.; Khalil, I.; Zomaya, A. Y.; Foufou, S.; and Bouras, A. 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3): 267–279.

Fang, Y.; Li, X.; Thomas, S.; and Zhu, X. 2023. ChatGPT as Data Augmentation for Compositional Generalization: A Case Study in Open Intent Detection. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, 13–33.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Guan, R.; Li, J.; Wang, S.; Tu, W.; Li, M.; Zhu, E.; Liu, X.; and Chen, P. 2025a. Multi-view Graph Clustering with Dual Relation Optimization for Remote Sensing Data. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7346–7355.

Guan, R.; Liu, T.; Tu, W.; Tang, C.; Luo, W.; and Liu, X. 2025b. Sampling Enhanced Contrastive Multi-View Remote Sensing Data Clustering with Long-Short Range Information Mining. *IEEE Transactions on Knowledge and Data Engineering*, 1–15.

Guan, R.; Tu, W.; Wang, S.; Liu, J.; Hu, D.; Tang, C.; Feng, Y.; Li, J.; Xiao, B.; and Liu, X. 2025c. Structure-Adaptive Multi-View Graph Clustering for Remote Sensing Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16933–16941.

Hakkani-Tür, D.; Ju, Y.-C.; Zweig, G.; and Tür, G. 2015. Clustering novel intents in a conversational interaction system with semantic parsing. In *INTERSPEECH*, 1854–1858.

Haponchyk, I.; Uva, A.; Yu, S.; Uryupina, O.; and Moschitti, A. 2018. Supervised clustering of questions into intents for dialog system applications. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2310–2321.

Hu, Z.; Zheng, W.; Zong, Y.; Wei, M.; Jiang, X.; and Shi, M. 2024. A Novel Decoupled Prototype Completion Network for Incomplete Multimodal Emotion Recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Kumar, R.; Patidar, M.; Varshney, V.; Vig, L.; and Shroff, G. 2022. Intent detection and discovery from user logs via deep semi-supervised contrastive clustering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1836–1853.

Lin, T.-E.; Xu, H.; and Zhang, H. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8360–8367.

Liu, J.; Liu, X.; Wang, S.; Wan, X.; Li, D.; Lu, K.; and He, K. 2025. Communication-efficient federated multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, J.; Liu, X.; Yang, Y.; Liao, Q.; and Xia, Y. 2023. Contrastive multi-view kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9552–9566.

Liu, S.; Liao, Q.; Wang, S.; Liu, X.; and Zhu, E. 2024a. Robust and consistent anchor graph learning for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 4207–4219.

Liu, S.; Zhang, J.; Wen, Y.; Yang, X.; Wang, S.; Zhang, Y.; Zhu, E.; Tang, C.; Zhao, L.; and Liu, X. 2024b. Sample-level cross-view similarity learning for incomplete multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 14017–14025.

Liu, Z.; Zhou, B.; Chu, D.; Sun, Y.; and Meng, L. 2024c. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 101: 101973.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Min, Q.; Qin, L.; Teng, Z.; Liu, X.; and Zhang, Y. 2021. Dialogue state induction using neural latent variable models. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 3845–3852.

Mou, Y.; He, K.; Wu, Y.; Zeng, Z.; Xu, H.; Jiang, H.; Wu, W.; and Xu, W. 2022. Disentangled knowledge transfer for OOD intent discovery with unified contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 46–53.

Mou, Y.; Song, X.; He, K.; Zeng, C.; Wang, P.; Wang, J.; Xian, Y.; and Xu, W. 2023. Decoupling Pseudo Label Disambiguation and Representation Learning for Generalized

Intent Discovery. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9661–9675.

Padmasundari, S. B.; and Bangalore, S. 2018. Intent discovery through unsupervised semantic text clustering. In *Proc. Interspeech*, volume 2018, 606–610.

Peng, L.; Ye, Y.; Liu, C.; Che, H.; Wang, F.; Yu, Z.; Wu, S.; and Wong, H.-S. 2025. SMART: Semantic Matching Contrastive Learning for Partially View-Aligned Clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Saha, T.; Upadhyaya, A.; Saha, S.; and Bhattacharyya, P. 2021. Towards sentiment and emotion aided multi-modal speech act classification in twitter. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5727–5737.

Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Patel, O. P.; Tiwari, A.; Er, M. J.; Ding, W.; and Lin, C.-T. 2017. A review of clustering techniques and developments. *Neurocomputing*, 267: 664–681.

Shi, W.; An, W.; Tian, F.; Zheng, Q.; Wang, Q.; and Chen, P. 2023. A Diffusion Weighted Graph Framework for New Intent Discovery. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8033–8042.

Song, X.; He, K.; Wang, P.; Dong, G.; Mou, Y.; Wang, J.; Xian, Y.; Cai, X.; and Xu, W. 2023. Large Language Models Meet Open-World Intent Discovery and Recognition: An Evaluation of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10291–10304.

Tian, J.; Wang, K.; Xu, X.; Cao, Z.; Shen, F.; and Shen, H. T. 2022. Multimodal disentanglement variational autoencoders for zero-shot cross-modal retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 960–969.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, 6558.

Ulmer, D.; Hardmeier, C.; and Frellsen, J. 2022. Deep-significance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vedula, N.; Lipka, N.; Maneriker, P.; and Parthasarathy, S. 2019. Towards open intent discovery for conversational text. *arXiv preprint arXiv:1904.08524*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.

Yang, D.; Huang, S.; Kuang, H.; Du, Y.; and Zhang, L. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM international conference on multimedia*, 1642–1651.

Zeng, Y.; Yan, W.; Mai, S.; and Hu, H. 2024. Disentanglement translation network for multimodal sentiment analysis. *Information Fusion*, 102: 102031.

Zhang, D.; Nan, F.; Wei, X.; Li, S.-W.; Zhu, H.; Mckeown, K.; Nallapati, R.; Arnold, A. O.; and Xiang, B. 2021a. Supporting Clustering with Contrastive Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5419–5430.

Zhang, H.; Xu, H.; Lin, T.-E.; and Lyu, R. 2021b. Discovering new intents with deep aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14365–14373.

Zhang, H.; Xu, H.; Long, F.; Wang, X.; and Gao, K. 2024. Unsupervised Multimodal Clustering for Semantics Discovery in Multimodal Utterances. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 18–35.

Zhang, H.; Xu, H.; Wang, X.; Long, F.; and Gao, K. 2023. A clustering framework for unsupervised and semi-supervised new intent discovery. *IEEE Transactions on Knowledge and Data Engineering*, 36(11): 5468–5481.

Zhang, H.; Xu, H.; Wang, X.; Zhou, Q.; Zhao, S.; and Teng, J. 2022a. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM international conference on multimedia*, 1688–1697.

Zhang, Y.; Zhang, H.; Zhan, L.-M.; Wu, X.-M.; and Lam, A. 2022b. New Intent Discovery with Pre-training and Contrastive Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 256–269.

Zhou, Y.; Quan, G.; and Qiu, X. 2023. A probabilistic framework for discovering new intents. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3771–3784.

Zhu, J.; Zou, X.; Liu, L.; Huang, Z.; Zhang, Y.; Tang, C.; and Dai, L.-R. 2025a. Trusted Mamba Contrastive Network for Multi-View Clustering. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhu, T.; Liu, Q.; Wang, F.; Tu, Z.; and Chen, M. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*.

Zhu, Z.; Zhou, P.; Li, Z.; Chen, K.; and Zhu, J. 2025b. Multi-label text classification with label attention aware and correlation aware contrastive learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 8420–8428.