InMu-Net: Advancing Multi-modal Intent Detection via Information Bottleneck and Multi-sensory Processing

Zhihong Zhu
Peking University
Beijing, China
zhihongzhu@stu.pku.edu.cn

Xuxin Cheng
Peking University
Beijing, China
chengxx@stu.pku.edu.cn

Zhaorun Chen
The University of Chicago
Chicago, USA
zhaorun@uchicago.edu

Yuyan Chen
Fudan University
Shanghai, China
chenyuyan21@m.fudan.edu.cn

Yunyan Zhang
Jarvis Research Center, Tencent
YouTu Lab
Shenzhen, China
yunyanzhang@tencent.com

Xian Wu*
Jarvis Research Center, Tencent
YouTu Lab
Shenzhen, China
kevinxwu@tencent.com

Yefeng Zheng
Medical Artificial Intelligence Lab,
Westlake University
Hangzhou, China
Jarvis Research Center, Tencent
YouTu Lab
Shenzhen, China
zhengyefeng@westlake.edu.cn

Bowen Xing
Beijing Key Laboratory of Knowledge
Engineering for Materials Science,
School of Computer and
Communication Engineering,
University of Science and Technology
Beijing
Beijing, China
bwxing714@gmail.com

ABSTRACT

Multi-modal intent detection (MID) aims to comprehend users' intentions through diverse modalities, which has received widespread attention in dialogue systems. Despite the promising advancements in complex fusion mechanisms or architecture designs, challenges remain due to: (1) various noise and redundancy in both visual and audio modalities and (2) long-tailed distributions of intent categories. In this paper, to tackle the above two issues, we propose InMu-Net, a simple yet effective framework for MID from the **In**formation bottleneck and **Mu**lti-sensory processing perspective. Our contributions lie in three aspects. First, we devise a *denoising bottleneck module* to filter out the intent-irrelevant information in the fused feature; Second, we introduce a *saliency preservation loss* to prevent the dropping of intent-relevant information; Ultimately, *kurtosis regulation* is introduced to maintain representation smoothness during the filtering process, mitigating the adverse impact of the long tail distribution. Comprehensive experiments on two MID benchmark datasets demonstrate the effectiveness of InMu-Net and its vital components. Impressively, a series of analyses reveal our denoising potential and robustness in low-resource, modality corruption, cross-architecture and cross-task scenarios.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681623

CCS CONCEPTS

• **Information systems → Multimedia information systems**; • **Computing methodologies → Discourse, dialogue and pragmatics**.

KEYWORDS

Multi-modal Intent Detection, Multi-modal Information Bottleneck, Multi-sensory Processing

1 INTRODUCTION

Intent detection (ID) aims to ascertain the objectives of users conveyed through their utterances, which serves as a crucial component of task-oriented dialogue systems [29, 35]. Prior studies have extensively researched ID and validated the significance of textual modality [24, 57]. However, beyond textual utterances, facial expressions and audio signals are also informative as they are often complementary and interact synergistically [48]. Therefore, multi-modal intent detection (MID) has attracted increasing research attention, which is more practical in real-world scenarios.

To effectively leverage the information from various modalities, numerous methods have been proposed for MID. Therein, Saha et al. [33], Zhang et al. [58] pioneered MID area by introducing tri-modal ID benchmarks with text, visual and audio information; Zhou et al. [59] proposed a token-level contrastive learning method with modality-aware prompting to facilitate modality fusion; Huang et al. [12] introduced a shallow-to-deep Transformer-based framework with ChatGPT-based data augmentation strategy to align different modality features, obtaining state-of-the-art (SOTA) results.

*Corresponding author

Figure 1: An example of redundancy in the visual modality in MIntRec [58] dataset. As indicated, the utterance is only relevant to the woman highlighted in gray, and the information like the distracting background is intent-irrelevant.

Despite promising advancements achieved, we discover that existing MID models still suffer from two main issues:

(1) **Noise and redundancy in visual and audio modalities.** As depicted in Figure 1, there exists plenty of redundancy and noise in the visual modality, which can impede the performance of multi-modal fusion. To this end, several studies proposed methods like gating mechanisms [19, 52] to filter noise and redundancy. However, an underexplored aspect is the potential of these fusion gates to filter vital information within the filtered modalities inadvertently.

(2) **Long-tailed distributions of intent categories.** Existing MID benchmarks [33, 58] exhibit a pronounced long-tailed distribution of intent categories as shown in Figure 2, where a few classes, *i.e.*, head classes, contain a major number of samples while the remaining classes, *i.e.*, tail classes, have only a small number of samples. A straightforward remedy is to rebalance the training dataset through weighted sampling. However, this is a suboptimal strategy that may be detrimental to the accuracy of the head classes [11, 47].

In this paper, we propose a new framework termed InMu-Net to address the existing issues jointly, drawing inspirations from **In**formation bottleneck [7, 46, 50] and **Mu**lti-sensory processing [26, 41]. Through three vital components, InMu-Net advances towards its denoising and redundancy reduction capabilities, alongside bolstering robustness across multiple categories. Specifically, ❶ we design a *denoising bottleneck module* to effectively reduce intent-irrelevant feature redundancy. ❷ we present a *saliency preservation loss*, which provides explicit supervision to maximize the intent-relevant information in the fused feature. ❸ we perform *kurtosis regulation* on both unimodal and multi-modal representations. In this manner, InMu-Net can diminish sensitivity towards tail intents, thus mitigating the adverse effects of long-tailed distribution. Quantitative experiments demonstrate that our InMu-Net significantly outperforms previous SOTA methods. Systematic analyses confirm the superiority of InMu-Net against distinct scenarios.

Contributions. In a nutshell, our contributions are three-fold:

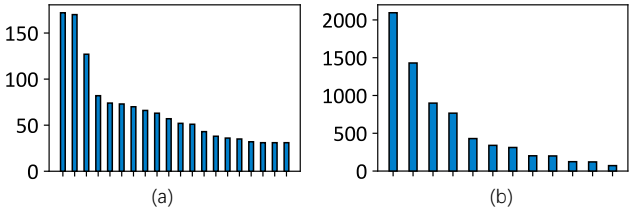

Figure 2: Distribution of intent categories on two MID training datasets: (a) MIntRec [58] and (b) MELD-DA [33].

- We present a new framework dubbed InMu-Net for MID, drawing inspirations from the information bottleneck and multi-sensory processing. To our best knowledge, this is the first attempt to bridge the information bottleneck and MID.
- We introduce three core modules in the proposed InMu-Net from the principled perspective, addressing significant issues of noise redundancy and the long-tail problem in MID.
- Extensive experiments including low-resource, modality corruption, cross-architecture, and cross-task scenarios demonstrate the generalizability and robustness of our InMu-Net.

2 PRELIMINARIES

2.1 Task Description

Formally, given a tri-modal input comprising text, visual and audio modalities, the multi-modal intent detection (MID) task can be conceptualized as a classification task that determines the intent label for the tri-modal input, which is expressed as follows:

$$y = f(X_t, X_v, X_a), \tag{1}$$

where $f(\cdot)$ represents the MID model; X_t, X_v, X_a denote the text, visual and audio input, respectively; $y \in Y = \{y_1, y_2, \ldots, y_K\}$ is the intent label associated with one of the K predefined intents.

2.2 Feature Encoding

To begin, we first encode the multi-modal sequential input X_m (where $m \in \{t, v, a\}$) into unit-length representations f_m. Specifically, we employ separate modality-specific Transformers [43] to capture the features of distinct modalities as follows:

$$f_m = \text{Transformer}_m(X_m; \theta_m). \tag{2}$$

in which θ_m represents the parameters of the Transformer_m.

2.3 Mutual Information

Mutual information [15] is a measure of the amount of information shared between random variables. Formally, it quantifies the statistical dependency of two random variables X and Y:

$$\mathcal{I}(X;Y) = \mathbb{E}_{p(X,Y)}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right], \tag{3}$$

where $p(X, Y)$ denotes the joint probability distribution between X and Y, while $p(X)$ and $p(Y)$ are their marginals.

2.4 Kurtosis Estimation

Kurtosis [4] is a statistical measure utilized to categorize the tail behavior of distributions. It is sensitive to rare events and is employed for analyzing distributions with "fatter tails". For *univariate*

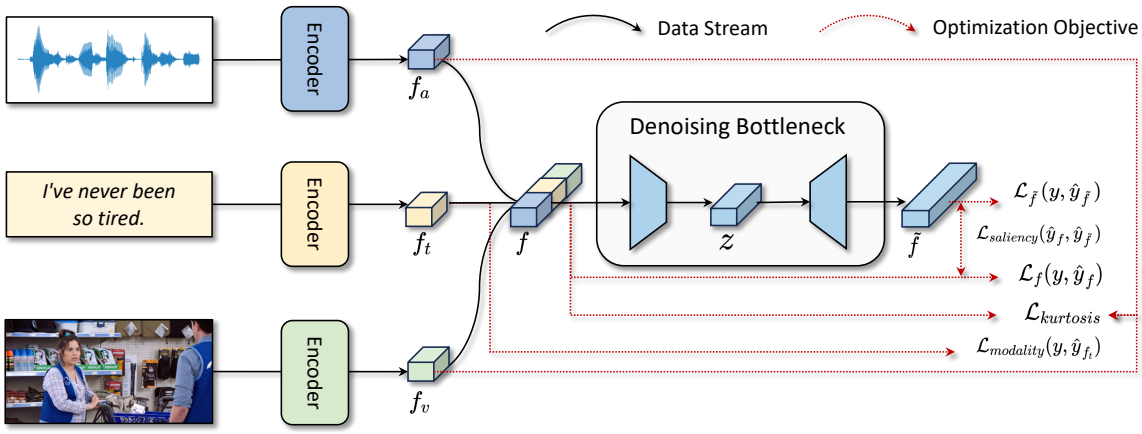

Figure 3: The architecture of InMu-Net. Mathematical symbols in the figure are consistent with the formulas in the paper.

variables, kurtosis is quantified as the standardized fourth moment:

$$\frac{\mathbb{E}\left[(X-\mu)^4\right]}{(\mathbb{E}\left[(X-\mu)^2\right])^2}. \tag{4}$$

It is commonly used to assess deviations from normalcy. Building upon this, Mardia [23] extended kurtosis to a *multivariate* context:

$$\mathbb{E}\left[((X-\mu)^\top \Sigma^{-1}(X-\mu))^2\right], \tag{5}$$

where X denotes a multi-dimensional random vector with μ, Σ representing the mean and covariance matrix of X, respectively.

3 METHODOLOGY

In this section, we detail our proposed framework (InMu-Net) for MID, the architecture of which is depicted in Figure 3. Concretely, we start by concatenating the features extracted from each modality (§2.2) to create an initial fused feature f, which retains comprehensive information from each modality. Then, We present the underlying motivations for our proposed supervision (§3.1). After that, we apply a denoising bottleneck module (§3.2) to perform information distillation, culminating in the refined feature $\tilde{f}$. Notably, we incorporate a saliency preservation loss to ensure no critical intent-relevant information is discarded (§3.3). Additionally, we introduce a kurtosis regulation loss (§3.4) inspired by neural synergy [3] to represent unimodal and multimodal features more effectively against long-tail distribution.

3.1 Supervision Motivation

As discussed above, we aim to eliminate redundancies and noise in the visual and audio modalities within MID. To this end, we resort to information bottleneck (IB) [22], which can find the concise and compressed representation of the input. By applying the IB principle, the model can learn to filter out noisy and redundant information that might otherwise interfere with prediction accuracy. Mathematically, consider the encoded fused feature $\tilde{f}$, derived from the direct fused feature f. Our objective is for $\tilde{f}$ to encapsulate only intent-relevant information while eliminating intent-irrelevant redundancies. Drawing on the principles of mutual information [7] in IB, the information between $\tilde{f}$ and f can be quantified as follows:

$$\begin{aligned} \mathcal{I}(f;\tilde{f}) = \mathcal{I}(\tilde{f};f) &= \mathbb{E}_{p(f,\tilde{f})}\left[\log \frac{p(f,\tilde{f})}{p(f)p(\tilde{f})}\right] \\ &= \iint p_{f,\tilde{f}}(f,\tilde{f}) \log \frac{p_{f,\tilde{f}}(f,\tilde{f})}{p_f(f)p_{\tilde{f}}(\tilde{f})} df d\tilde{f}. \end{aligned} \tag{6}$$

This expression can be further expanded by introducing the target variable y and applying the chain rule as follows:

$$\begin{aligned} \mathcal{I}(f;\tilde{f}) &= \iiint p(f,\tilde{f}|y)p(y) \log \frac{p(f,\tilde{f}|y)p(y)}{p(f)p(\tilde{f})} \, df \, d\tilde{f} \, dy \\ &= \iiint p(f,\tilde{f},y) \log \frac{p(f,\tilde{f}|y)}{p(f|y)p(\tilde{f}|y)} \, df \, d\tilde{f} \, dy \\ &+ \iint p(\tilde{f},y) \log \frac{p(\tilde{f}|y)}{p(\tilde{f})} \, d\tilde{f} \, dy \\ &= \mathcal{I}(f;\tilde{f}|y) + \mathcal{I}(\tilde{f};y), \end{aligned} \tag{7}$$

which distinguishes the intent-irrelevant information $\mathcal{I}(f;\tilde{f}|y)$ from the intent-relevant information $\mathcal{I}(\tilde{f};y)$. Thus, the goals of IB are to ❶ minimize the $\mathcal{I}(f;\tilde{f}|y)$ while ❷ maximize the $\mathcal{I}(\tilde{f};y)$. These dual goals form the objective of IB, which are defined as:

$$\min_{\tilde{f}} \mathcal{L}_{IB} = \mathcal{I}(f;\tilde{f}|y) - \gamma \mathcal{I}(\tilde{f};y), \tag{8}$$

in which γ is a scalar that determines the weight of the minimal information constraint $\mathcal{I}(\tilde{f};y)$ during optimization.

However, conducting a min-max game as formulated in Eq. (8) is challenging [25, 28] due to the well-documented difficulty in estimating mutual information in high-dimensional spaces [27]. To address this issue, we introduce a denoising bottleneck to achieve goal ❶ and saliency preservation to achieve goal ❷. Additionally, we implement kurtosis regulation to ensure smooth representations across both unimodal and multimodal contexts, thereby mitigating the negative effects associated with tailed intents.

3.2 Denoising Bottleneck

Considering three encoded features $f_{\{t,v,a\}}$ from distinct modalities (§2.2) in MID, we first generate an initial fused feature $f = [f_t, f_v, f_a] \in \mathbb{R}^n$, where $[\cdot,\cdot]$ denotes concatenation operation. Subsequently, we suggest automatic denoising at the feature level. To elaborate, the denoising bottleneck module comprises two linear projection layers along with dropout and ReLU activation functions:

$$f \xrightarrow[\text{Dropout+ReLU}]{\text{Linear Projection}} z \xrightarrow[\text{Dropout+ReLU}]{\text{Linear Projection}} \tilde{f}, \tag{9}$$

where z is of dimension $p < n$, while the final feature representation $\tilde{f}$ retains n dimensions. The re-projection of z back to the same dimensions as f serves two purposes: ❶ It facilitates feature-level supervision, enabling the learning of a more intent-relevant feature $\tilde{f}$ as discussed in the subsequent subsection. ❷ Empirically, we observed that the feature-level supervision proves more effective when $\tilde{f}$ and the initial feature f are aligned.

Remark 1. *There are several applications of the IB within the multimodal community. [9] explored the application of IB at different stages of fusion, while [50] implemented IB at the sequence level. Our method offers several advantages: ❶ simplicity in fusion approach; ❷ flexibility in handling variable-length sequence inputs. Additionally, direct comparisons are provided in the experiments (cf. Table 8).*

3.3 Saliency Preservation

The denoising bottleneck constrains information flow across modalities to filter out redundancy and noise. However, it might inadvertently lead to the loss of vital information as well [9]. Inspired by the success of mutual information in applications within the computer vision community [6, 20, 39], we introduce saliency preservation, which is designed to explicitly maximize intent-relevant information $\mathcal{I}(\tilde{f};y)$ in Eq.(8), ensuring that the process of information reduction does not compromise the quality of crucial information.

Theorem 1. *To elucidate, maximizing $\mathcal{I}(\tilde{f};y)$ can be interpreted as minimizing the difference in the mutual information between the original and the denoised features for y, formally expressed as:*

$$\min \mathcal{I}(f;y) - \mathcal{I}(\tilde{f};y) \Longleftrightarrow \min H(y|f) - H(y|\tilde{f}), \tag{10}$$

where $H(y|f)$ is defined as the conditional entropy:

$$H(y|f) := -\int p(f) d_f \int p(y|f) \log p(y|f) d_y. \tag{11}$$

Building on this theoretical foundation, we further validate the effectiveness of $\tilde{f}$ through the following corollary:

Corollary 1. *If the KL-divergence [16] between the predicted distributions of the fused feature f and the denoised $\tilde{f}$ equals to 0, then $\tilde{f}$ is sufficient for y as well i.e.,*

$$D_{KL}\left[p(y|f)||p(y|\tilde{f})\right] = 0 \Longrightarrow \mathcal{I}(y;f) - \mathcal{I}(y;\tilde{f}) = 0, \tag{12}$$

where $p(y|f), p(y|\tilde{f})$ represent the predicted distributions under f and $\tilde{f}$, respectively; and $D_{KL}(\cdot)$ denotes the KL-divergence.

To operationalize this theory, we finally frame our saliency preservation loss as follows:

$$\mathcal{L}_{saliency} = D_{KL}\left[p(y|f)||p(y|\tilde{f})\right]. \tag{13}$$

3.4 Kurtosis Regulation

In multi-sensory neural processing, around 20% of the neurons account for 80% of the information propagation in cortical circuits [26]. Several researches by [5, 40] demonstrated that multi-modal computation tends to concentrate in such local cortical clusters and found significantly lower kurtosis in such clusters. Besides, low kurtosis helps alleviate sensitivity to rare events, as discussed in §2.4. These all suggest that individual unimodal and multi-modal representations in MID should exhibit low levels of kurtosis. To maintain this characteristic, we regulate the multivariate kurtosis by plugging in standard estimators for the mean and covariates:

$$\mathcal{L}_{kurtosis} = \frac{1}{N}\sum_{j=1}^{N}\left[((f_j - \bar{f})^\top S^{-1}(f_j - \bar{f}))^2\right], \tag{14}$$

where N denotes the number of features, f_j represents samples from features, including both unimodal features like $f_{\{t,v,a\}}$ and the direct fused feature f. $\bar{f}$ denotes the empirical mean feature $\bar{f} = \frac{\sum_{j=1}^{n} f_j}{N}$ and S signifies the empirical covariance matrix:

$$S = \frac{\sum_{i=1}^{N}(f_j - \bar{f})(f_j - \bar{f})^\top}{N-1}. \tag{15}$$

Since $\tilde{f}$ is derived from f, we do not perform regularization on $\tilde{f}$. Note that the covariance matrix is computed via a decaying moving average over a window of multiple batches to produce smoother estimates before the inversion operation.

3.5 Overall Objective

Eventually, the overall loss function in our InMu-Net is:

$$\begin{aligned}\mathcal{L} = &\underbrace{\mathcal{L}_f(y,\hat{y}_f) + \mathcal{L}_{modality}(y,\hat{y}_{f_t}) + \mathcal{L}_{\tilde{f}}(y,\hat{y}_{\tilde{f}})}_{\text{Foundational Supervision}} \\ &+ \alpha\mathcal{L}_{saliency}(\hat{y}_f,\hat{y}_{\tilde{f}}) + \beta\mathcal{L}_{kurtosis},\end{aligned} \tag{16}$$

where $\mathcal{L}_f$ and $\mathcal{L}_{\tilde{f}}$ are respective losses to supervise the direct fused feature and denoised feature, and $\mathcal{L}_{modality}$ supervises core textual modality encoder since textual information dominates across modalities [13, 38]. α and β are trade-off hyper-parameters. $\hat{y}_f$ and $\hat{y}_{\tilde{f}}$ are the classifier results from the direct fused feature and denoised feature, respectively. Note that the first three losses are directly supervised by y and serve as the foundational supervision of the overall framework in the form of cross-entropy losses.

During inference, the denoised feature $\tilde{f}$ is employed to determine the ultimate intent. Consequently, InMu-Net serves as an augmentation during the training phase, with only the *denoising bottleneck module* invoked during inference, given that simple Linear and Dropout contribute minimally to inference latency.

4 EXPERIMENTS

4.1 Datasets and Metrics

We conduct experiments on two benchmarks to evaluate the proposed InMu-Net: ❶ **MIntRec** [58],[1] which is a fine-grained dataset for multi-modal intent recognition. It comprises 2,224 high-quality samples across text, visual and audio modalities, distributed among

[1] https://github.com/thuiar/MIntRec/tree/main

Model	MIntRec				MELD-DA			
	ACC	wF1	wP	R	ACC	wF1	wP	R
MAG-BERT [30]	72.65	72.16	72.53	69.28	60.63	59.36	59.80	50.01
MulT [42]	72.52	72.31	72.85	69.24	60.36	59.01	59.44	49.93
MISA [10]	72.29	72.38	73.48	69.24	59.98	58.52	59.28	48.75
TCL-MAP [59]	73.62	73.31	73.72	70.50	61.75	59.77	60.33	50.14
SDIF-DA* [12]	73.90	73.93	73.96	71.61	61.31	58.01	60.93	49.96
InMu-Net (Ours)	**76.05**†	**75.96**†	**76.18**†	**73.93**†	**63.78**†	**61.64**†	**63.40**†	**52.31**†

Table 1: Experimental results on two MID datasets. Best scores are in bold and second-best scores are in underlined. Results with $*$ are obtained by re-implemented, while others are taken from the corresponding published paper. † denotes the significant paired t-tests of our InMu-Net over the baseline models at p-value < 0.05.

Model	Common					Long-tail				
	Complain	Inform	Praise	Apologise	Thank	Agree	Flaunt	Oppose	Ask for help	Joke
MAG-BERT	67.65	71.00	86.03	97.76	96.52	91.60	47.09	33.97	64.44	37.54
MulT	65.48	70.85	84.72	97.93	96.83	92.23	48.91	34.68	69.12	33.95
MISA	63.91	70.18	86.63	97.78	**98.03**	92.05	46.44	36.15	67.57	38.74
TCL-MAP	**68.70**	**72.80**	87.20	97.70	97.00	93.10	50.80	35.90	66.40	29.00
SDIF-DA*	67.76	71.24	87.67	98.11	97.96	92.31	44.44	30.02	67.02	45.56
InMu-Net	67.76	71.09	**89.25**	**98.63**	**98.03**	**94.42**	**56.52**	**41.38**	**69.58**	**55.38**
	↑ −0.94	↑ −1.71	↑ 1.58	↑ 0.52	↑ 0.00	↑ 1.32	↑ 5.72	↑ 5.23	↑ 0.46	↑ 9.82
w/o $\mathcal{L}_{kurtosis}$	67.40	70.79	89.01	98.24	97.25	93.38	48.42	36.80	66.83	48.69
	↓ 0.36	↓ 0.30	↓ 0.24	↓ 0.39	↓ 0.78	↓ 1.04	↓ 8.10	↓ 4.58	↓ 2.75	↓ 6.69
Human♮	80.08	79.69	93.44	96.15	96.90	87.21	78.10	69.04	88.54	72.22

Table 2: F1-score comparison of common and long-tail subsets on MIntRec. ↑ *Number* denotes the improvement our method achieves in the current category compared to the best baseline; ↓ *Number* represents the decrease in model performance across different intent categories after removing the proposed Kurtosis Regulation. Results with ♮ are taken from Zhou et al. [59].

20 intent categories. The dataset is divided into 1,334 training samples, 445 validation samples, and 445 testing samples. ❷ **MELD-DA** [33],[2] which is a large-scale dataset designed for dialogue act classification. It includes 9,988 multi-modal samples annotated across 12 common dialogue act labels, with a split of 6,991 training samples, 999 validation samples and 1,998 testing samples.

Following previous works, we employ accuracy (ACC), weighted F1-score (wF1), weighted precision (wP), and recall (R) as evaluation metrics to assess the proposed InMu-Net framework. To account for category imbalances, the wF1 and wP metrics are calculated as weighted averages, with weights corresponding to the sample counts in each category. Unless specified otherwise, higher values indicate better performance across all metrics in this work.

4.2 Implementation Details

For a fair comparison, we follow Huang et al. [12], Zhang et al. [58] to adopt bert-base-uncased [14] and wav2vec2-base-960h [1] from Huggingface Library [49] to extract text and audio features and Faster R-CNN [31] from Torchvision Library to extract visual features. AdamW [21] is utilized as the optimizer with a learning rate searched from $[1e^{-6}, 3e^{-5}]$. The batch size is set as 16 for training and 8 for validation/testing. For hyper-parameter α and β, we test them in the range from 0.2 to 1.0 on the validation set and choose the best-performing one to the test set, respectively. Paired t-test is performed to test the significance of performance improvement with a default significance level of 0.05. All experiments are conducted on one single NVIDIA GeForce RTX 3090. The results reported in all experiments are averages of 5 random runs.

[2]https://github.com/thuiar/TCL-MAP

4.3 Main Results

We compare the proposed InMu-Net framework with a series of competitive MID baselines, including: MAG-BERT [30], MulT [42], MISA [10], TCL-MAP [59] and SDIF-DA [12]. The main results on two benchamark datasets are reported in Table 1, from which we have the following observations:

❶ Our proposed InMu-Net consistently outperforms all baselines on both MIntRec and MELD-DA datasets and achieves a new SOTA performance. Specifically, on MIntRec dataset, it overpasses the previous SOTA model SDIF-DA by 2.91% and 2.75% on ACC and wF1, respectively; on MELD-DA dataset, it overpasses TCL-MAP by 3.29% and 3.13% on ACC and wF1, respectively. This verifies the effectiveness of InMu-Net in the MID task. Furthermore, the significance tests of InMu-Net over the baseline models show that our InMu-Net significantly outperforms the baseline models (the results of p-value on all evaluation metrics are less than 0.05).

❷ Notably, the gains on MELD-DA are more pronounced. We suppose the reason is that MELD-DA is more challenging, whose data redundancy problem is more serious, involving complex scenes

Setting	$\mathcal{L}_f$	DB	$\mathcal{L}_{\tilde{f}}$	$\mathcal{L}_{saliency}$	$\mathcal{L}_{modality}$	$\mathcal{L}_{kurtosis}$	MIntRec				MELD-DA			
							ACC	wF1	wP	R	ACC	wF1	wP	R
(a)	✓	-	-	-	-	-	69.68	68.40	68.84	67.81	58.25	55.03	57.26	47.98
(b)	✓	-	-	-	t	✓	71.49	70.31	70.75	69.99	60.51	58.28	60.42	50.16
(c)	✓	✓	✓	-	t	✓	74.12	72.87	73.22	72.18	62.59	59.50	61.67	51.35
(d)	✓	✓	✓	✓	t	-	74.40	73.13	73.50	72.46	62.94	59.69	61.91	51.67
(e)	✓	✓	✓	✓	-	✓	72.81	71.48	71.89	70.88	60.98	57.92	60.04	50.34
(f)	✓	✓	✓	✓	a	✓	73.18	71.82	72.26	71.25	61.27	58.17	60.44	50.62
(g)	✓	✓	✓	✓	v	✓	73.07	71.75	72.21	71.10	61.12	58.11	60.38	50.53
(h)	✓	✓	✓	✓	t	✓	**76.05**	**75.96**	**76.18**	**73.93**	**63.78**	**61.64**	**63.40**	**52.31**

Table 3: Ablation studies. "DB" is short for denoising bottleneck. "t, a, v" denotes textual, audio, and visual modality, respectively.

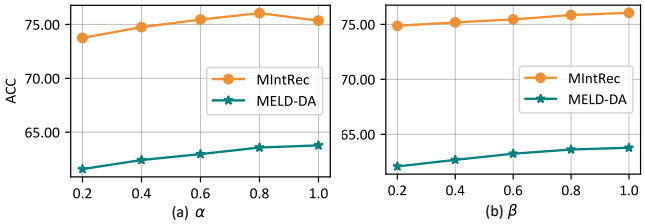

Figure 4: Hyper-parameter analyses. Effect of the trade-off hyper-parameter (a) α and (b) β in Eq.(16).

Model	Perplexity (↓)
TCL-MAP [59]	2.55
SDIF-DA [12]	2.48
InMu-Net	1.82

Table 4: Comparison with SOTA baselines on perplexity. ↓ denotes lower is better.

and overlapping characters. The proposed InMu-Net excels in denoising and preserving vital information, coupled with enhanced robustness via kurtosis regulation, resulting in superior performance over strong baseline models TCL-MAP and SDIF-DA.

❸ To clearly demonstrate how our method addresses the long-tail distribution issue, we selected the top five and bottom five intents based on their occurrence frequencies in the training set. These are referred to as the "Common" and "Long-tail" subsets, respectively, and their F1 scores are presented in Table 2. We find our InMu-Net achieves superior performance in 8 out of the 10 intent categories, particularly showing notable improvements in the "Long-tail" intents. Thanks to the kurtosis regularization, InMu-Net can adaptively mitigate the adverse effects of the long-tail distribution, achieving robust performance.

4.4 Ablation Study

We perform a set of ablation studies to understand the necessity of the different designs and strategies in the proposed InMu-Net. From the results in Table 3, we can obtain the following takeways:

Denoising Bottlneck. The denoising bottleneck module coupled with saliency preservation loss is first removed to not perform information filtering. By comparing the row setting (b) to setting (h) in Table 3, we can conclude that the decreased performance implies that the proposed denoising bottleneck actually contributes to eliminating redundancy and noise within the modalities.

Saliency Preservation. There is a consistent performance degradation on both datasets (comparing setting (c) to setting (h)) when the denoising bottleneck lacks saliency preservation supervision. A plausible deduction is that the supervision mitigates the loss of valuable information due to intent-irrelevant feature filtering.

Kurtosis Regualion. ❶ Here, we remove the kurtosis regulation to verify its effectiveness (comparing setting (d) to setting (h)). The poor results show that appropriate regulating multi-modal features can boost model performance. ❷ Furthermore, when comparing the data in Table 2 without the $\mathcal{L}_{kurtosis}$ term (*w/o* $\mathcal{L}_{kurtosis}$), a noticeable decline in model performance is observed. This further demonstrates that the proposed kurtosis regularization can effectively alleviate the sensitivity to tail intents and mitigate long-tail distribution issues, thereby enabling more robust predictions.

Centre Modality. As mentioned in §3.5, results by text-centric supervision tend to perform better as low information intensity and high redundancy in other modalities. Thus, we evaluate results based on no, audio and visual modality supervision (setting (e), (f) and (g)). We observe an obvious decline in performance when audio, visual, or no modality is used as the central supervision.

4.5 Hyper-parameter Analysis

Since Eq.(16) encompasses multiple loss components, we delved into the influence of the core elements as shown in Figure 4. Specifically, we select the best-performing model based on the validation set and then evaluate it on the test set: ❶ The parameter α indicates the importance of $\mathcal{L}_{saliency}$. We evaluate the scale range setting $\alpha \in [0.2, 1.0]$ as shown in Figure 4(a). We find that accuracy is improved and saturated with 0.8 and 1.0 on MIntRec and MELD-DA, respectively. Thus, we set $\alpha = 0.8$ for MIntRec and 1.0 for MELD-DA in practice. ❷ The parameter β signifies the extent of involvement of $\mathcal{L}_{kurtosis}$. Our observations indicate a relative insensitivity to parameter selection, as the incorporation of kurtosis regulation generally yields gains for MID, albeit to varying degrees. As a result, we adopt $\beta = 1.0$ to achieve the best performance.

4.6 Perplexity Evaluation

Since the denoising module plays an important role in the proposed approach, we provide more insight analysis about it. Specifically, we evaluated the perplexity for the prediction of the golden label across two state-of-the-art baselines (*i.e.*, TCL-MAP and SDIF-DA) and our proposed InMu-Net on the MIntRec test set. From the results in

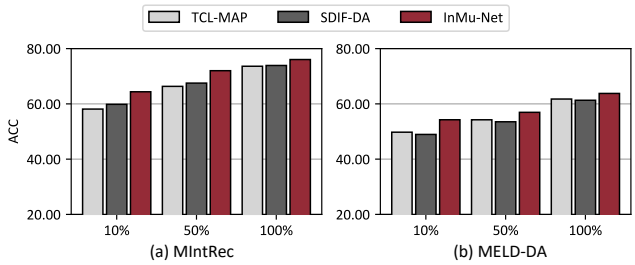

Figure 5: Performances on low-resource settings.

Modality			MIntRec		MELD-DA	
t	*a*	*v*	ACC	wF1	ACC	wF1
-	✓	✓	58.48	57.27	47.43	44.96
✓	✓	-	68.76	67.35	57.02	54.13
✓	-	✓	70.14	68.98	58.54	55.47
✓	✓	✓	**76.05**	**75.96**	**63.78**	**61.64**

Table 5: Results on modality corruption.

Table 4, we observe that both TCL-MAP and SDIF-DA exhibit perplexity around 2.5, whereas our method demonstrates a significant reduction (-0.7). This compellingly illustrates the effectiveness of our denoising bottleneck module in accurately predicting intents.

4.7 Low-resource Settings

To investigate the effect and robustness of InMu-Net in low-resource scenarios, we conduct experiments using different limited training sizes following Huang et al. [12]. From the results in Figure 5, we observe that InMu-Net consistently outperforms SOTA baselines, especially when the resource is quite limited (10%). This indicates that InMu-Net can sustain denoising capabilities within resource-constrained scenarios and distill intent-relevant information from multi-modal representations, thus boosting performance.

4.8 Modality Corruption

We further assess the model performance by removing one modality at a time. From the results in Table 5, we find that ❶ the tri-modal combination yields the highest performance, indicating that our InMu-Net can learn complementary information from different modalities. ❷ the performance drops sharply when the textual modality is removed. We attribute this to the fact that textual modality has higher information density compared to redundant audio and visual modalities. This underscores two critical insights: First, eliminating noise and redundancy to enhance the information density of visual and audio modalities is crucial during fusion. Second, text-central fusion results may help boost performance in MID.

4.9 Generalizability Analysis

Analysis of Cross-architecture Scenario. To evaluate the generalizability of our proposed InMu-Net, we conduct preliminary experiments on two representative baselines MuIT and SDIF-DA. To be specific, we retain the baseline loss function for a fair comparison, integrating only our three vital components and corresponding supervisions on the fused feature. The results depicted in Table 6

Model	MIntRec		MELD-DA	
	ACC	wF1	ACC	wF1
MulT* [42]	72.23	71.98	60.27	58.95
with InMu-Net	**74.01**	**73.85**	**61.53**	**59.75**
SDIF-DA* [12]	73.90	73.93	61.31	58.01
with InMu-Net	**75.25**	**75.08**	**62.94**	**60.38**

Table 6: Results on cross-architecture scenarios.

Model	Latency/Inference Time per Sample	Speedup
TCL-MAP [59]	25.8ms	1.0x
SDIF-DA [12]	25.4ms	1.0x
InMu-Net	25.5ms	1.0x

Table 7: Comparison on computation (inference) latency.

indicate that baselines augmented with InMu-Net outperform their original counterparts. This verifies our work's contribution is orthogonal to theirs, considering changing the architecture of InMu-Net for better multi-modal fusion is still a promising avenue.

Analysis of Cross-task Scenario. We further conduct comparison experiments on multi-modal sentiment analysis (MSA) to evaluate the proposed InMu-Net and the results are reported in Table 8. It can be observed that InMu-Net achieves competitive performance compared with MSA baselines, which verifies that InMu-Net can generalize the denoising ability to different tasks. Note that MMIM performs hierarchical mutual information maximization for each modality, which is deeply integrated with the overall framework, and proposes a parameterized method to approximate the true value. Whereas, even without careful hyper-parameters tuning, our proposed InMu-Net outperforms MMIM on 4 out of 6 metrics and achieves similar performance on the remaining two metrics.

4.10 Computation Efficiency

As the proposed method is applicable on real-world applications, we conducted a preliminary latency comparison among the two most advanced baseline models and our method, with the results reported in the Table 7. We find that our model's inference speed surpasses that of TCL-MAP and maintains a speedup on par with current SOTA methods. This can be attributed to the fact that our method functions as an augmentation during the training phase, with only the denoising bottleneck module being activated during inference. Given that simple Linear and Dropout layers contribute minimally to the inference latency, this ensures efficient performance.

4.11 Visualization

To qualitatively demonstrate how our proposed InMu-Net filters noise and redundancy while capturing precise intent-relevant information, we provide GradCAM-CAM [34] visualizations of InMu-Net and SOTA baseline SDIF-DA. From Figure 6 case (a), it can be seen that while both models SDIF-DA and InMu-Net successfully identify the critical term "*gift*" in the visual modality. However, the capture range of SDIF-DA is broader and more diffuse, including irrelevant background details. In contrast, our proposed InMu-Net focuses narrowly and precisely on the gift on the table, demonstrating its ability to filter out extraneous information and pinpoint

Model	MOSI [55]			MOSEI [56]		
	MAE(↓)	Corr(↑)	Acc-7(↑)	MAE(↓)	Corr(↑)	Acc-7(↑)
ICCN [36]	0.860	0.710	39.0	0.565	0.713	51.6
MISA [10]	0.783	0.761	42.3	0.555	0.756	52.2
Self-MM [54]	0.712	0.795	45.8	0.529	0.767	53.5
MMIM [9]	0.700	0.800	46.7	0.526	0.772	54.2
DBF [50]	**0.693**	**0.801**	44.8	0.523	0.772	54.2
InMu-Net (Ours)	0.694	0.798	**46.9**	**0.520**	**0.774**	**54.5**

Table 8: Results of the experiments on cross-task scenarios. ↓ denotes lower is better. ↑ denotes higher is better.

Figure 6: Comparison of Grad-CAM [34] visualizations between SOTA baseline SDIF-DA and our InMu-Net.

intent-relevant details with high accuracy. A similar situation can be observed in Figure 6 case (b), where the baseline model SDIF-DA exhibits excessive redundant attention, while our InMu-Net accurately captures the woman speaking in the figure.

5 RELATED WORK

Multi-modal Intent Detection. Multi-modal intent detection (MID) is a significant task for understanding human language in task-oriented dialogue systems [2, 60–62]. Compared to text-only intent detection, which ascertains the objectives of users conveyed through their utterances, MID integrates facial expressions and audio signals to fully leverage the complementary and interactive information provided by diverse modalities. A series of models [12, 33, 58, 59] have been proposed and made promising progress. Therein, Zhou et al. [59] introduced a token-level contrastive learning coupled with modality-aware prompting to improve modality fusion. Concurrently, Huang et al. [12] developed a Transformer-based [44, 45] framework that progresses from shallow to deep interactions, complemented by ChatGPT-based data augmentation techniques to align features across modalities.

Our work do not focus on intricate fusion mechanisms [37] or architectural intricacies, our approach emphasizes the perspective of multi-modal information and data distribution.

Information Bottleneck. The InfoMax proposed by Linsker [18] seeks to maximize the mutual information between feature and model output. Along this way, there have been many works that explore optimal ways for mutual information estimation. Han et al. [9] built up a hierarchical mutual information maximization guided model for multi-modal sentiment analysis. Wu et al. [50] focused on video-based sentiment analysis and performed contrastive learning to achieve mutual information maximization. Despite these advancements, these theories in MID remain under-studied.

In contrast to existing works, our InMu-Net devises a denoising bottleneck and a saliency preservation loss to precisely filter intent-irrelevant information and keep intent-relevant information in an adaptive manner, which obviates the reliance on heuristic or greedy feature selection methods [17, 50] and coarse-grained gate filtering mechanism [51, 53]. Furthermore, the proposed feature-level denoising offers advantages in handling variable-length inputs.

Multi-sensory Processing. In multi-sensory processing resarch area, different sensory modalities are processed individually and then combined in various multimodal convergence zones, including cortical and subcortical regions [8]. Several research such as Faber et al. [5], Timme et al. [40] demonstrated that multi-modal computation tends to concentrate in such local cortical clusters and found significantly lower kurtosis in such clusters.

In this work, we resort to multivariate kurtosis [32] to alleviate the long-tail distribution problem in MID. The most straightforward remedy to this problem is to rebalance the training dataset through weighted sampling. However, it is a suboptimal strategy that may be detrimental to the accuracy of the head classes [47]. Innovatively, we treat the fused multi-modal feature as cortical clusters, controlling peaking to reduce sensitivity to tailed intents.

6 CONCLUSION

In this paper, we proposed InMu-Net, a new framework from the information bottleneck and multi-sensory processing perspectives, to tackle modality redundancy and long-tailed distribution of labels in MID jointly. InMu-Net maximizes the intent-relevant information in fused multi-modal features by the proposed *denosing bottleneck* and minimizes the intent-irrelevant information by the proposed *saliency preservation* loss. Moreover, *kurtosis regulation* is introduced to reduce the negative impact of long-tail distributions. Extensive experiments and analyses on two MID benchmarks show the superiority of our proposed framework.

REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[2] Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. 2024. AutoPRM: Automating Procedural Supervision for Multi-Step Reasoning via Controllable Question Decomposition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1346–1362. https://aclanthology.org/2024.naacl-long.73

[3] Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

[4] Lawrence T DeCarlo. 1997. On the meaning and use of kurtosis. *Psychological methods* 2, 3 (1997), 292.

[5] Samantha P Faber, Nicholas M Timme, John M Beggs, and Ehren L Newman. 2019. Computation is concentrated in rich clubs of local cortical networks. *Network Neuroscience* 3, 2 (2019), 384–404.

[6] Yingying Fang, Shuang Wu, Sheng Zhang, Chaoyan Huang, Tieyong Zeng, Xiaodan Xing, Simon Walsh, and Guang Yang. 2024. Dynamic Multimodal Information Bottleneck for Multimodality Classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 7696–7706.

[7] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2019. Learning Robust Representations via Multi-View Information Bottleneck. In *International Conference on Learning Representations*.

[8] Asif A Ghazanfar and Charles E Schroeder. 2006. Is neocortex essentially multisensory? *Trends in cognitive sciences* 10, 6 (2006), 278–285.

[9] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9180–9192.

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

[11] Xinting Hu, Yi Jiang, Kaihua Tang, Jingyuan Chen, Chunyan Miao, and Hanwang Zhang. 2020. Learning to segment the tail. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14045–14054.

[12] Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. 2024. SDIF-DA: A Shallow-to-Deep Interaction Framework with Data Augmentation for Multi-modal Intent Detection. *ICASSP* (2024).

[13] Mengzhao Jia, Can Xie, and Liqiang Jing. 2023. Debiasing Multimodal Sarcasm Detection with Contrastive Learning. *arXiv preprint arXiv:2312.10493* (2023).

[14] Jacob Devlin Kenton, Ming-Wei Chang, and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.

[15] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E* 69, 6 (2004), 066138.

[16] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[17] Fan Li, Xu Si, Shisong Tang, Dingmin Wang, Kunyan Han, Bing Han, Guorui Zhou, Yang Song, and Hechang Chen. 2024. Contextual Distillation Model for Diversified Recommendation. *arXiv preprint arXiv:2406.09021* (2024).

[18] Ralph Linsker. 1988. Self-organization in a perceptual network. *Computer* 21, 3 (1988), 105–117.

[19] Nayu Liu, Xian Sun, Hongfeng Yu, Fanglong Yao, Guangluan Xu, and Kun Fu. 2022. Abstractive summarization for video: A revisit in multistage fusion network with forget gate. *IEEE Transactions on Multimedia* (2022).

[20] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang Wang. 2022. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11006–11016.

[21] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[22] Sijie Mai, Ying Zeng, and Haifeng Hu. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia* (2022).

[23] Kanti V Mardia. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 3 (1970), 519–530.

[24] Jie Mei, Yufan Wang, Xinhui Tu, Ming Dong, and Tingting He. 2023. Incorporating BERT with probability-aware gate for spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 826–834.

[25] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory* 56, 11 (2010), 5847–5861.

[26] Sunny Nigam, Masanori Shimono, Shinya Ito, Fang-Chin Yeh, Nicholas Timme, Maxym Myroshnychenko, Christopher C Lapish, Zachary Tosi, Pawel Hottowy, Wesley C Smith, et al. 2016. Rich-club organization in effective connectivity among cortical neurons. *Journal of Neuroscience* 36, 3 (2016), 670–684.

[27] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. 2019. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems* 32 (2019).

[28] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*. PMLR, 5171–5180.

[29] Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A Survey on Spoken Language Understanding: Recent Advances and New Frontiers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4577–4584. Survey Track.

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

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[32] Fernando E Rosas, Pedro AM Mediano, Michael Gastpar, and Henrik J Jensen. 2019. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E* 100, 3 (2019), 032305.

[33] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4361–4372.

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[35] Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 4661–4676.

[36] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8992–8999.

[37] Shisong Tang, Qing Li, Xiaoteng Ma, Ci Gao, Dingmin Wang, Yong Jiang, Qian Ma, Aoyang Zhang, and Hechang Chen. 2022. Knowledge-based Temporal Fusion Network for Interpretable Online Video Popularity Prediction. In *Proceedings of the ACM Web Conference 2022*. 2879–2887.

[38] Shisong Tang, Qing Li, Dingmin Wang, Ci Gao, Wentao Xiao, Dan Zhao, Yong Jiang, Qian Ma, and Aoyang Zhang. 2023. Counterfactual Video Recommendation for Duration Debiasing. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4894–4903.

[39] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. 2021. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1522–1531.

[40] Nicholas M Timme, Shinya Ito, Maxym Myroshnychenko, Sunny Nigam, Masanori Shimono, Fang-Chin Yeh, Pawel Hottowy, Alan M Litke, and John M Beggs. 2016. High-degree neurons feed cortical computations. *PLoS computational biology* 12, 5 (2016), e1004858.

[41] Christopher Town. 2007. Multi-sensory and multi-modal fusion for sentient computing. *International Journal of Computer Vision* 71 (2007), 235–253.

[42] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6558–6569.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[44] Zhongwei Wan, Xinjian Wu, Yu Zhang, Yi Xin, Chaofan Tao, Zhihong Zhu, Xin Wang, Siqi Luo, Jing Xiong, and Mi Zhang. 2024. D2O: Dynamic Discriminative Operations for Efficient Generative Inference of Large Language Models. *arXiv preprint arXiv:2406.13035* (2024).

[45] Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. 2024. LOOK-M: Look-Once Optimization in KV Cache for Efficient Multimodal Long-Context Inference. *arXiv preprint arXiv:2406.18139* (2024).

[46] Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu. 2021. Multi-view information-bottleneck representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10085–10092.

[47] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. 2017. Learning to model the tail. *Advances in neural information processing systems* 30 (2017).

[48] Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. 2023. Tackling Modality Heterogeneity with Multi-View Calibration Network for Multimodal Sentiment Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5240–5252.

[49] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[50] Shaoxiang Wu, Damai Dai, Ziwei Qin, Tianyu Liu, Binghuai Lin, Yunbo Cao, and Zhifang Sui. 2023. Denoising Bottleneck with Mutual Information Maximization for Video Multimodal Fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2231–2243.

[51] Yifeng Xie, Zhihong Zhu, Xuan Lu, Zhiqi Huang, and Haoran Xiong. 2024. InfoEnh: Towards Multimodal Sentiment Analysis via Information Bottleneck Filter and Optimal Transport Alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 9073–9083. https://aclanthology.org/2024.lrec-main.795

[52] Zihui Xue and Radu Marculescu. 2023. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2574–2583.

[53] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3995–4007.

[54] 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 *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10790–10797.

[55] 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* (2016).

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

[57] Feng Zhang, Wei Chen, Fei Ding, and Tengjiao Wang. 2023. Dual Class Knowledge Propagation Network for Multi-label Few-shot Intent Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 8605–8618.

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

[59] 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. *AAAI* (2024).

[60] Zhihong Zhu, Xuxin Cheng, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Code-Switching Can be Better Aligners: Advancing Cross-Lingual SLU through Representation-Level and Prediction-Level Alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

[61] Zhihong Zhu, Xuxin Cheng, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Dance with Labels: Dual-Heterogeneous Label Graph Interaction for Multi-intent Spoken Language Understanding. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (Merida, Mexico) *(WSDM '24)*. Association for Computing Machinery, New York, NY, USA, 1022–1031. https://doi.org/10.1145/3616855.3635782

[62] Zhihong Zhu, Xuxin Cheng, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2024. Aligner^2: Enhancing Joint Multiple Intent Detection and Slot Filling via Adjustive and Forced Cross-Task Alignment. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 17 (Mar. 2024), 19777–19785. https://doi.org/10.1609/aaai.v38i17.29952