Dual-oriented Disentangled Network with Counterfactual Intervention for Multimodal Intent Detection

Zhanpeng Chen, Zhihong Zhu, Xianwei Zhuang, Zhiqi Huang, Yuexian Zou[*]
ADSPLAB, School of ECE, Peking University, China
{troychen927,zhihongzhu,xwzhuang}@stu.pku.edu.cn
{zhiqihuang,zouyx}@pku.edu.cn

Abstract

Multimodal intent detection leverages diverse modalities for a comprehensive understanding of user intentions in real-world scenarios, playing a critical role in modern task-oriented dialogue systems. While existing methods have made progress in modal alignment and fusion, they overlook two vital limitations: (I) *Close entanglement of multimodal semantics with modal structures;* (II) *Insufficient learning of the causal effects of semantic and modality-specific information on final predictions in end-to-end training.* To address these limitations, we introduce the **Du**al-**o**riented **D**isentangled **N**etwork with Counterfactual Intervention (**DuoDN**). DuoDN consists of a Dual-oriented Disentangled Encoder that decouples semantics- and modality-oriented representations, and a Counterfactual Intervention Module that uses causal inference to understand causal effects by injecting confounders. Experiments on three benchmark datasets demonstrate DuoDN's superiority over existing methods, with extensive analysis validating its advantages.

1 Introduction

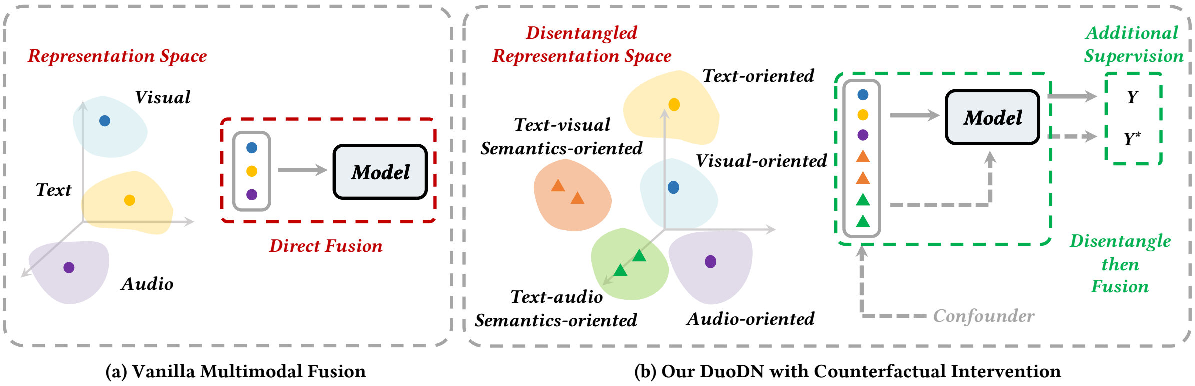
With the rise of intelligent technology, task-oriented dialogue systems are rapidly advancing and proving their potential in practical applications like health consulting, financial services, and home automation. A crucial component of these systems is natural language understanding (NLU), which captures the semantics of user queries to enhance conversational interactions (Qin et al., 2021). After converting spoken words to text via speech recognition, the system needs to understand the user's intent to respond accurately. Recent years have seen significant advancements in text-based intent detection (Qin et al., 2019; Zhu et al., 2023; Xing and Tsang, 2022). However, as modern dialogue systems increasingly use multimodal interactions, it is essential to study multimodal intent detection for more practical applications.

Recently, research has focused on bi-modal intent detection using textual guidance (Gonzaga et al., 2021; Agrawal et al., 2022; Huang et al., 2020), which is a step forward but still falls short for real-world natural language understanding. To address this, Zhang et al. (2022) introduced MIntRec, the first tri-modal intent detection benchmark, incorporating text, video, and audio. Building on this, Huang et al. (2023) proposed the Shallow-to-Deep Interaction Framework with Data Augmentation (SDIF-DA), which aligns multimodal features for better fusion and uses ChatGPT to generate additional utterances, tackling data scarcity. Zhou et al. (2023) introduced a token-level contrastive learning method with modality-aware prompting (TCL-MAP), creating an optimal multimodal semantic space to refine the text modality. However, these approaches either overlook the rich, complex semantic information from different modalities or neglect the causal effects on final predictions from semantic and modality-specific information.

Addressing the first issue, simply fusing multimodal information is challenging due to the persistent gaps between heterogeneous modalities, which are often complex and interdependent. These interdependencies are frequently overlooked, resulting in inadequate task modeling. Regarding the second issue, optimizing the model through end-to-end learning often leads to suboptimal outcomes, weakening the focus on semantic and modality-specific information. In summary, existing methods face two major limitations in complex scenarios: (I) *Multimodal semantic information is deeply intertwined with modality-specific structures*; (II) *The causal effects of semantic and modality-specific information on final predictions are not properly learned in end-to-end training*.

[*]Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17554–17567
November 12-16, 2024 ©2024 Association for Computational Linguistics



(a) Vanilla Multimodal Fusion

(b) Our DuoDN with Counterfactual Intervention

Figure 1: Illustrations of different multimodal fusion paradigms. (a) simply fuses the multimodal representations from the feature extractors. (b) learns multimodal representations through semantics-oriented and modality-oriented subspaces. These features are later utilized for fusion and optimized by different supervisions.

To address the first limitation, we employ disentangled representation learning. Drawing inspiration from advances in domain adaptation (Bousmalis et al., 2016), we aim to learn two distinct types of representations for each modality. The process begins with encoders that encode the tri-modal inputs. Then, a dual-oriented disentangled encoder decouples the multimodal representations into semantics-oriented and modality-oriented factors. On one hand, the semantics-oriented representations aim to reduce modality gaps. Despite originating from different sources, multimodal information often shares common motives. The semantic-oriented encoder captures these commonalities as aligned projections in a shared subspace. To enhance semantic consistency, we apply semantic-level contrastive learning on the disentangled semantics-oriented features, bringing similar semantic features from different modalities closer together. On the other hand, multimodal signals embody both commonalities and distinctive characteristics. Each modality's unique traits can be useful for predicting the speakers' intentions. Therefore, the modality-oriented representations complement the shared semantic features and promote comprehensive fusion learning.

To address the second limitation, we use causal inference to analyze the pure causal effects of semantic- and modality-specific information on final predictions. Causal inference, widely applied in video understanding (Qi et al., 2020, 2021) and audio-visual video question answering (Li et al., 2022, 2023), reveals implicit causal relationships among variables and enhances model generalization. End-to-end training often lacks directed supervision towards significant features, leading to suboptimal outputs. Inspired by causal inference, we incorporate a counterfactual intervention module as additional supervision. This module highlights the semantics- and modality-oriented representations by injecting confounders, enabling the model to understand causal effects. We introduce the indirect effect by maximizing the difference between the original output and a counterfactual output, altered only by changes in the semantic- or modality-oriented representations. This helps the model perceive their functions. Both the original and intervened outputs are used to optimize supervision, resulting in a comprehensive understanding of causal effects and improving the model's performance and robustness.

Our DuoDN has been shown to outperform previous state-of-the-art methods on three benchmark datasets: MIntRec (Zhang et al., 2022), MIntRec 2.0 (Zhang et al., 2024), and MELD-DA (Saha et al., 2020), demonstrating its effectiveness. Specifically, it exhibits an excellent understanding of semantics, achieving superior performance on MIntRec 2.0, both with and without out-of-scope samples. Additionally, ablation analysis confirms that our proposed methods of disentangled representation learning and counterfactual intervention are effective, contributing to advancements in related research.

To summarize, our major contributions are as follows: (I) We propose a dual-oriented disentangled network with counterfactual intervention for multimodal intent detection, effectively disentangling and utilizing modality- and semantic-specific information. (II) To the best of our knowledge, we are the first to introduce causal inference to determine the causal effects of semantic- and modality-specific information on the final predictions in multimodal intent detection. (III) Extensive experiments on three benchmarks show that DuoDN significantly outperforms previous methods. Further analysis verifies the advantages of our model.

2 Related Work

2.1 Intent Analysis

Recognizing user intent is crucial for NLU, which uses text to determine intent for better conversations. While text-based intent recognition works well for specific tasks (Qin et al., 2019; Zhang et al., 2021; Coucke et al., 2018; Casanueva et al., 2020), it doesn't handle real-world multimodal language, including emotions, attitudes, and behaviors. Adding non-verbal signals like expressions, body movements, and tone can enhance understanding and user experience.

Recently, multimodal language understanding has advanced with new datasets, boosting research and applications (Yagcioglu et al., 2018; Yu et al., 2020; Zadeh et al., 2018; Xie et al., 2024). MIntRec and MIntRec 2.0 have inspired innovative multimodal intent detection methods like SDIF-DA and TCL-MAP. In this paper, we focus on separating and linking different modalities, improving both regular and out-of-scope multimodal intent analysis.

2.2 Disentangled Representation Learning

Disentangled representation learning aims to identify the underlying factors of observable data (Bengio et al., 2013) and has been widely used in computer vision and natural language processing (Chen et al., 2021; Bao et al., 2019). However, these methods are still limited in multimodal intent detection, which requires combining complementary information from different modalities. To address this, we propose an integrated dual-oriented disentangled encoder with semantic-level contrastive learning to better align semantic feature spaces.

2.3 Causal Learning

In recent years, counterfactual thinking and causal reasoning have gained popularity in visual explanations (Goyal et al., 2019; Wang and Vasconcelos, 2020; Yi et al., 2019) and video understanding (Qi et al., 2021, 2019). For example, Rao et al. (2021) used counterfactual training to address spatial attention bias in fine-grained image recognition, while Niu et al. (2021) reduced language bias in visual question answering by separating the direct language effect from the total causal effect. Unlike these approaches, we focus on the perspective of enhancing semantics-oriented and modality-oriented representations to optimize final predictions.

3 Method

3.1 Preliminaries

Task Formulation Specifically, given a trimodal input including $\boldsymbol{x}_T$, $\boldsymbol{x}_V$ and $\boldsymbol{x}_A$, corresponding respectively to textual, visual, and auditory modalities, the multimodal intent detection can be seen as a classification task to decide the intent label $\boldsymbol{y}$ of the inputs.

Feature Extraction To obtain the feature extraction of each modality, we adopt BERT (Devlin et al., 2018), WavLM (Chen et al., 2022), and Swin Transformer (Liu et al., 2021) following the approach suggested by Zhang et al. (2024). The text embedding $\boldsymbol{T}$ and audio embedding $\boldsymbol{A}$ are directly derived from respective encoders. Regarding the video features, we extract the video keyframes and process each frame through the Swin Transformer, which is pre-trained on ImageNet (Deng et al., 2009). We apply RoIAlign (He et al., 2017) to feature maps of keyframes using annotated RoIs to convert them to fixed sizes. Finally, we generate overall RoI feature embeddings $\boldsymbol{V}$ by average pooling the feature maps.

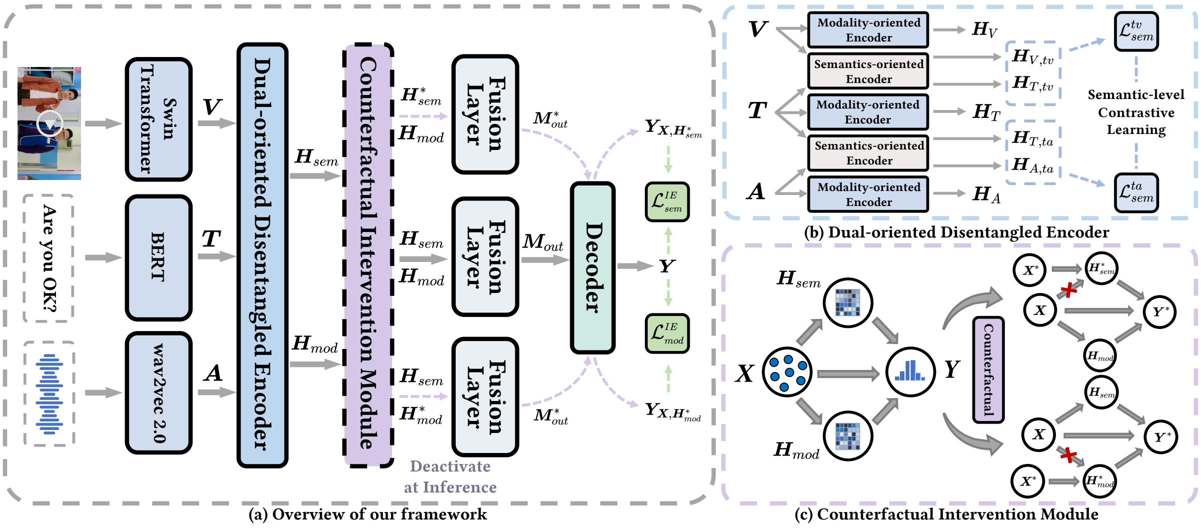
3.2 Dual-oriented Disentangled Encoder

Semantics-Oriented Encoder The semantics-oriented encoder captures the shared semantics between modalities, which means the semantics-oriented representation from the same encoder should be close in the feature space. According to Hazarika et al. (2020) and Guo et al. (2022), various modalities contribute differently to pattern recognition, with the text modality being particularly dominant. Merely combining these modalities can result in ambiguous or confusing semantics, thereby significantly limiting overall performance. Therefore, we utilize two MLPs to learn the latent representations of text-video and text-audio pair:

$$\boldsymbol{H}_{V,tv} = MLP_{sem,tv}(\boldsymbol{V}), \boldsymbol{H}_{T,tv} = MLP_{sem,tv}(\boldsymbol{T}), \quad (1)$$

$$\boldsymbol{H}_{A,ta} = MLP_{sem,ta}(\boldsymbol{A}), \boldsymbol{H}_{T,ta} = MLP_{sem,ta}(\boldsymbol{T}), \quad (2)$$

where $\boldsymbol{H}_{V,tv}$, $\boldsymbol{H}_{T,tv}$, $\boldsymbol{H}_{A,ta}$ and $\boldsymbol{H}_{T,ta}$ refer to the semantic representation of video, text in the text-video pair, audio, and text in the text-audio pair,



(a) Overview of our framework

(b) Dual-oriented Disentangled Encoder

(c) Counterfactual Intervention Module

Figure 2: The illustration of our proposed framework DuoDN. To endow the model with the ability to pursue the causal effect, the counterfactual intervention module (CIM) is activated during training and deactivated during inference. In the inference phase, $\boldsymbol{H}_{sem}$ and $\boldsymbol{H}_{mod}$ are directly fed into the fusion layer. Note that $\boldsymbol{H}_{sem} \in \{\boldsymbol{H}_{T,tv}, \boldsymbol{H}_{V,tv}, \boldsymbol{H}_{T,ta}, \boldsymbol{H}_{A,ta}\}$ and $\boldsymbol{H}_{mod} \in \{\boldsymbol{H}_T, \boldsymbol{H}_V, \boldsymbol{H}_A\}$.

respectively. In order to enhance semantic consistency between modalities, we apply contrastive learning (Chen et al., 2020; Oord et al., 2018) to learn the semantics-oriented representations, which can be formulated as follows,

$$\mathcal{L}_{sem}^{tv} = -\log \frac{exp(s(\boldsymbol{H}_{V,tv}^i, \boldsymbol{H}_{T,tv}^i)/\tau)}{\sum_{m\in\{T,V\}}\sum_{j\neq i} exp(s(\boldsymbol{H}_{m,tv}^i, \boldsymbol{H}_{m,tv}^j)/\tau)}, \tag{3}$$

$$\mathcal{L}_{sem}^{ta} = -\log \frac{exp(s(\boldsymbol{H}_{A,ta}^i, \boldsymbol{H}_{T,ta}^i)/\tau)}{\sum_{m\in\{T,A\}}\sum_{j\neq i} exp(s(\boldsymbol{H}_{m,ta}^i, \boldsymbol{H}_{m,ta}^j)/\tau)}, \tag{4}$$

where the matched pair serves as a positive example and the different pairs from the same batch serve as a negative example. $s(\cdot,\cdot)$ means cosine similarity function and τ is a temperature parameter. With the contrastive loss incorporated, matched pairs are robustly aligned while unrelated examples are pushed away to promote uniformity in the representation space (Wang and Isola, 2020).

Modality-Oriented Encoder To capture unique information coupled tightly in different sensory channels, we disentangle the multimodal representations and extract their modality-oriented representations by employing distinct MLP as follows,

$$\boldsymbol{H}_m = MLP_m(\boldsymbol{m}),\ m \in \{T, V, A\}. \tag{5}$$

After disentangling the multimodal representations into two solely encoded parts, we learn various hidden explanatory factors behind observable data (Bengio et al., 2013), which is fundamental for further counterfactual learning and fusion layers.

3.3 Counterfactual Intervention Module

To better understand the influence that the semantic-level and modality-level information brings, we focus on causal inference to endow our model with the ability to pursue the causal effect. As shown in Figure 2, we consider the message-passing process as a Structural Causal Model (Pearl et al., 2016; Pearl and Mackenzie, 2018). After obtaining the dual-oriented representations from the dual-oriented disentangled encoder, we observe that the indirect path $\boldsymbol{X} \rightarrow \boldsymbol{H}_{sem} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{X} \rightarrow \boldsymbol{H}_{mod} \rightarrow \boldsymbol{Y}$ do not gain enough attention because of the end-to-end training fashion. Drawing inspiration from Pearl et al. (2016), intervening in confounding factors can illuminate the significance of specific reasoning features. Thus, we carry out additional supervision to highlight the semantic-oriented and modality-oriented representation by introducing the Indirect Effect (Pearl, 2022) as follows,

$$IE_{sem}(\boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{Y}) = \mathbb{E}(\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}_{sem}}) - \mathbb{E}(\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}^*_{sem}}), \tag{6}$$

$$IE_{mod}(\boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{Y}) = \mathbb{E}(\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}_{mod}}) - \mathbb{E}(\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}^*_{mod}}), \tag{7}$$

where $\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}_{sem}}$ and $\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}_{mod}}$ refer to the original outputs of feeding forward the original input $\boldsymbol{X} = \{\boldsymbol{x}_T, \boldsymbol{x}_V, \boldsymbol{x}_A\}$. We intervene the message passing process by substituting the input $\boldsymbol{X}$ with confounder $\boldsymbol{X}^*$ using Gaussian distribution sampling:

$$\boldsymbol{X}^* = \boldsymbol{X}_\sigma \boldsymbol{W} + \boldsymbol{X}_\mu, \tag{8}$$

where $\boldsymbol{W}$ is the standard random vector with the same dimension of $\boldsymbol{X}$. In order to guarantee the gradient backpropagation, we use the reparameterization trick (Kingma and Welling, 2013) to learn the mean $\boldsymbol{X}_\mu$ and standard deviation $\boldsymbol{X}_\sigma$ instead of sampling discretely. Note that we utilize different random vector to sample $\boldsymbol{X}^*$ for different indirect path $\boldsymbol{X} \to \boldsymbol{H}_{sem} \to \boldsymbol{Y}$ and $\boldsymbol{X} \to \boldsymbol{H}_{mod} \to \boldsymbol{Y}$. Therefore, $\boldsymbol{H}^*_{sem}$ and $\boldsymbol{H}^*_{mod}$ represent the intervened dual-oriented representations after manually intervening the inputs. Obviously, $\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}^*_{sem}}$ and $\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}^*_{mod}}$ are impossible because $\boldsymbol{H}^*_{sem}$ and $\boldsymbol{H}^*_{mod}$ come from the confounder $\boldsymbol{X}^*$. So modification from $\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}_{sem}}/\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}_{mod}}$ to $\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}^*_{sem}}/\boldsymbol{Y}_{\boldsymbol{X},\boldsymbol{H}^*_{mod}}$ is equivalent to changing $\boldsymbol{H}_{sem}/\boldsymbol{H}_{mod}$ solely, which reveals the pure impact introduced by the confounder.

After the confounder is injected, the counterfactual classification results are expected to be worse than the original results because the intervened variables $\boldsymbol{H}^*_{sem}$ and $\boldsymbol{H}^*_{mod}$ commonly do not match with inputs $\boldsymbol{X}$. Since our goal is to instruct the model to increase the gap between the original output and the counterfactual one, we maximize IE_{sem} and IE_{mod} on the prediction of the correct class by minimizing the cross-entropy loss as follows,

$$\mathcal{L}^{IE}_{sem} = \mathcal{L}_{ce}(IE_{sem}),\ \mathcal{L}^{IE}_{mod} = \mathcal{L}_{ce}(IE_{mod}). \quad (9)$$

Consequently, the model is constrained to increase the difference between the outputs derived from $\boldsymbol{X}$ and $\boldsymbol{X}^*$ under the supervision of $\mathcal{L}^{IE}_{sem}$ and $\mathcal{L}^{IE}_{mod}$, leading to better training of the semantic-oriented and modality-oriented representation.

3.4 Fusion, Interaction and Optimization

Fusion Layer To acquire a cohesive joint multimodal representation, we apply Transformer (Vaswani et al., 2017) to perform the semantic-level fusion and modality-level interaction. The self-attention is employed to model the intra-modal interactions while the semantics-oriented representations of text are used to guide common information extraction through cross-attention. Firstly, we conduct the semantic-level fusion as follows,

$$\boldsymbol{Q}_{tv} = \boldsymbol{H}_{T,tv}\boldsymbol{W}^q_{T,tv}, \quad (10)$$

$$\boldsymbol{Q}_{ta} = \boldsymbol{H}_{T,ta}\boldsymbol{W}^q_{T,ta}, \quad (11)$$

$$\boldsymbol{K}_{tv} = [\boldsymbol{H}_{T,tv}\boldsymbol{W}^k_{T,tv} \oplus \boldsymbol{H}_{V,tv}\boldsymbol{W}^k_{V,tv}], \quad (12)$$

$$\boldsymbol{K}_{ta} = [\boldsymbol{H}_{T,ta}\boldsymbol{W}^k_{T,ta} \oplus \boldsymbol{H}_{A,ta}\boldsymbol{W}^k_{A,ta}], \quad (13)$$

$$\boldsymbol{V}_{tv} = [\boldsymbol{H}_{T,tv}\boldsymbol{W}^v_{T,tv} \oplus \boldsymbol{H}_{V,tv}\boldsymbol{W}^v_{V,tv}], \quad (14)$$

$$\boldsymbol{V}_{ta} = [\boldsymbol{H}_{T,ta}\boldsymbol{W}^v_{T,ta} \oplus \boldsymbol{H}_{A,ta}\boldsymbol{W}^v_{A,ta}], \quad (15)$$

$$\bar{\boldsymbol{H}}_{TV} = Attention(\boldsymbol{Q}_{tv}, \boldsymbol{K}_{tv}, \boldsymbol{V}_{tv}), \quad (16)$$

$$\bar{\boldsymbol{H}}_{TA} = Attention(\boldsymbol{Q}_{ta}, \boldsymbol{K}_{ta}, \boldsymbol{V}_{ta}), \quad (17)$$

where $\oplus$ denotes concatenation. Subsequently, the modality-level interaction is conducted with $\boldsymbol{Q}_m = \boldsymbol{H}_m\boldsymbol{W}^q_m$, $\boldsymbol{K}_m = \boldsymbol{H}_m\boldsymbol{W}^k_m$, and $\boldsymbol{V}_m = \boldsymbol{H}_m\boldsymbol{W}^v_m$ as follows,

$$\bar{\boldsymbol{H}}_m = Attention(\boldsymbol{Q}_m, \boldsymbol{K}_m, \boldsymbol{V}_m),\ m \in \{T, V, A\}, \quad (18)$$

where $\boldsymbol{W}^{q/k/v}_{T,tv}$, $\boldsymbol{W}^{k/v}_{V,tv}$, $\boldsymbol{W}^{q/k/v}_{T,ta}$, $\boldsymbol{W}^{k/v}_{A,ta}$, and $\boldsymbol{W}^{q/k/v}_m$ are learnable parameters of linear transformations. By applying the fusion and the interaction, we ensure that the text-video and text-audio pair are aware of their paired representations from different modalities and subspaces. This helps each representation gather useful information from other representations that are complementary and synergistic, resulting in an overall effective orientation. Finally, we take the output and construct a joint-vector $\boldsymbol{M}_{out} = [\bar{\boldsymbol{H}}_{TV} \oplus \bar{\boldsymbol{H}}_{TA} \oplus \bar{\boldsymbol{H}}_T \oplus \bar{\boldsymbol{H}}_V \oplus \bar{\boldsymbol{H}}_A]$. We project into the label space through the decoder:

$$\hat{\boldsymbol{Y}} = MLP(\boldsymbol{M}_{out}). \quad (19)$$

Optimization The total loss of our method is defined as follows,

$$\mathcal{L} = \mathcal{L}^{tv}_{sem} + \mathcal{L}^{ta}_{sem} + \mathcal{L}^{IE}_{sem} + \mathcal{L}^{IE}_{mod} + \mathcal{L}_{cls}, \quad (20)$$

where $\mathcal{L}_{cls} = \mathcal{L}_{ce}(\hat{\boldsymbol{Y}})$ is the overall classification loss optimized by cross-entropy.

4 Experiments

4.1 Experimental Setup

Datasets & Evaluation Metrics We conduct experiments on three challenging multimodal datasets, which are MIntRec (Zhang et al., 2022), MELD-DA (Saha et al., 2020), and MIntRec 2.0 (Zhang et al., 2024), to evaluate our proposed framework. More details about the benchmark datasets can be found in Appendix A.

Following previous work (Zhou et al., 2023; Zhang et al., 2024), we adopt four metrics to evaluate the model performance on in-scope classification: accuracy (*ACC*), weighted F1-score (*WF1*), weighted precision (*WP*) and recall (*R*). To evaluate out-of-scope classification performance, we utilize accuracy (*ACC*), macro F1-score over all classes (*F1*), in-scope classes (*F1-IS*), and the out-of-scope class (*F1-OOS*).

Methods	MIntRec				MELD-DA			
	ACC(%)	WF1(%)	WP(%)	R(%)	ACC(%)	WF1(%)	WP(%)	R(%)
MAG-BERT♣	72.65	72.16	72.53	69.28	60.63	59.36	59.80	50.01
MulT♣	72.52	72.31	72.85	69.24	60.36	59.01	59.44	49.93
MISA♣	72.29	72.38	73.48	69.24	59.98	58.52	59.28	48.75
SDIF-DA*	73.90	73.93	73.96	71.61	61.31	58.01	60.93	49.96
TCL-MAP	73.62	73.31	73.72	70.50	61.75	59.77	60.33	50.14
DuoDN (ours)	**75.28**†	**75.09**†	**75.80**†	**71.77**	**62.86**†	**60.90**†	**62.13**†	**51.63**†

Table 1: Main results on MIntRec and MELD-DA. ♣ denotes the results from Zhou et al. (2023). * are from our re-implementation. The best results are in **bold** and the second best ones are underlined. † denotes our model significantly outperforms baselines with $p < 0.05$ under t-test.

Methods	In-scope Classification				In-scope + Out-of-scope Classification			
	ACC(%)	WF1(%)	WP(%)	R(%)	F1-IS(%)	ACC(%)	F1-OOS(%)	F1(%)
w/o Out-of-scope samples								
TEXT♠	59.30	58.01	58.85	51.31	43.37	43.24	30.40	42.96
MAG-BERT♠	60.58	59.68	59.98	55.10	46.48	44.80	34.03	46.08
MulT♠	60.66	59.55	60.12	53.77	45.65	46.14	38.57	45.42
DuoDN (ours)	**61.89**†	**61.12**†	**60.39**	**56.44**†	**47.89**†	**48.11**†	**40.06**†	**47.69**†
w/ Out-of-scope samples								
TEXT♠	59.99	58.62	58.65	52.11	45.83	55.61	61.54	46.34
MAG-BERT♠	60.12	59.11	58.83	53.79	47.52	56.20	62.47	48.00
MulT♠	60.18	58.82	59.38	52.56	46.88	56.00	61.66	47.35
DuoDN (ours)	**61.43**†	**60.77**†	**59.41**	**55.01**†	**48.99**†	**57.76**†	**63.95**†	**49.33**†

Table 2: Main results on MIntRec 2.0. ♠ denotes the results from Zhang et al. (2024).

Implementation Details When benchmarking on MIntRec and MELD-DA, we employ pre-trained *bert-base-uncased*, *WavLM* and *swin_b* as feature extractors for text, audio, and video. To ensure a fair comparison, we substitute *bert-base-uncased* with *bert-large-uncased* while benchmarking on MIntRec 2.0, as per Zhang et al. (2024). To optimize the total loss, we employ AdamW (Loshchilov and Hutter, 2017) and conduct experiments on 8 NVIDIA GeForce RTX 3090 GPUs. For all experiments, the results are obtained by averaging the scores over five runs with different random seeds.

4.2 Comparative Baselines

We perform a comprehensive comparative study against DuoDN by considering baselines listed below: (1) MAG-BERT (Rahman et al., 2020); (2) MulT (Tsai et al., 2019); (3) MISA (Hazarika et al., 2020); (4) SDIF-DA (Huang et al., 2023); (5) TCL-MAP (Zhou et al., 2023).

4.3 Main Results

Results on MIntRec and MELD-DA Comparing DuoDN with state-of-the-art baselines, the results are presented in Table 1. As indicated by the results, our method outperforms all the baselines across all four metrics on both datasets. On one hand, we observe that on MIntRec DuoDN demonstrates improvement of 1.38% on ACC, 1.16% on wF1 and 1.84% on wP, which indicates an outstanding ability of our method to effectively leverage multimodal information. Our strategy of disentangling and utilizing the modality-specific information and multi-modal semantic information, while introducing causal inference to determine the causal effects, is proved to be efficacious. On the other hand, our proposed model also achieves significant improvements on MELD-DA, which is challenging due to ambiguous dialogue actions such as 'Backchannel' and 'Acknowledge'. The performance on MELD-DA not only shows the robustness of DuoDN but also affirms its capability to recognize ambiguous intents such as dialogue actions.

Results on MIntRec 2.0 The performance of our DuoDN on the MIntRec 2.0 dataset is presented in Table 2. For single-turn dialogue experiments, we conduct two settings: training without out-of-scope samples (*w/o OOS*) and with out-of-scope samples (*w/ OOS*) following Zhang et al. (2024). Our DuoDN outperforms the other methods in all metrics both with and without out-of-scope samples, which implies its effectiveness and robustness. Though DuoDN suffers a slight decline in in-scope classification evaluation, the great

Model	MIntRec				MELD-DA			
	ACC (%)	WF1 (%)	WP (%)	R (%)	ACC (%)	WF1 (%)	WP (%)	R (%)
w/o single modality								
w/o Visual	73.11	72.87	73.39	69.66	61.32	59.45	60.89	50.34
w/o Audio	73.23	73.54	74.32	70.04	60.87	59.11	60.32	50.03
w/ all modalities								
w/o CIM	73.48	73.63	74.74	70.83	61.86	59.69	61.53	50.96
w/o SL	73.93	74.00	75.12	70.91	61.76	60.12	61.13	51.07
w/o Duo	72.58	72.51	73.42	68.63	60.46	58.33	60.21	49.02
w/o CIM & SL	72.13	71.80	72.50	68.80	60.40	58.69	59.78	49.63
DuoDN	**75.28**	**75.09**	**75.80**	**71.77**	**62.86**	**60.90**	**62.13**	**51.63**

Table 3: Ablation experiments of modules in DuoDN on MIntRec and MELD-DA. *CIM* denotes the counterfactual intervention module, *SL* denotes the semantic-level contrastive learning, and *Duo* denotes the dual-oriented disentangled encoder.

improvements in out-of-scope detection (F1-OOS and ACC) still demonstrate its superior capability to take full advantage of multimodal information. By disentangling the modality representations in a dual-oriented way and enhancing with the counterfactual intervention, our proposed method naturally earns the ability to effectively deal with complex scenarios such as the out-of-scope classification. Additionally, the accuracy (ACC) and F1 scores are generally lower when out-of-scope samples are included in the in-scope classification, indicating the challenge out-of-scope samples present to real-world multimodal intent detection.

4.4 Ablation Study

To better understand how each part of DuoDN affects its performance, we conduct ablation experiments, and the results are shown in Table 3.

Effect of Modality We examined the contribution of each modality by removing the audio and visual modalities separately. Since the text modality is central to the disentangling and fusion process and contains less noise and redundancy than the other two modalities, its removal would cause a significant performance decline, which we did not detail here. The results clearly show that the multimodal combination yields the best performance, indicating that the model learns complementary features. Without this case, the tri-modal combination would not perform better than the bi-modal variants, such as text-visual DuoDN.

Effect of Dual-oriented Disentangled Encoder Replacing the dual-oriented disentangled encoder with an MLP for each modality results in a noticeable decline in WF1, with decreases of 3.44% on MIntRec and 4.22% on MELD-DA. Duo provides performance gains whether used alone or with other modules, highlighting its fundamental role in disentangling multimodal inputs and capturing semantic and modality-specific information.

Effect of Counterfactual Intervention Module Our core contribution, the counterfactual intervention module (CIM), enhances training by leveraging hidden states. As shown in Table 3, CIM improves all metrics by approximately 2%. Without CIM, the performance drops sharply even when Duo is activated, confirming the importance of enhancing the learning of semantics-oriented and modality-oriented representations. CIM helps the model learn a structure closer to the ground truth.

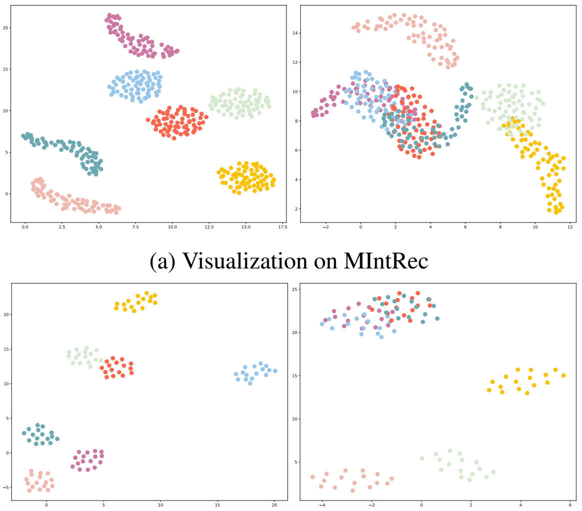
Effect of Semantic-level Contrastive Learning Removing the semantic-level contrastive learning losses from the total loss leads to a performance decline, as shown in Table 3. This change is attributed to insufficient alignment between different modalities in the semantic space. When the contrastive loss is active, the semantics between modalities align more consistently, enhancing DuoDN's ability to capture multimodal semantic information.

4.5 Fine-grained Analysis on Intent Taxonomies

As depicted in Figure 4, we visualize the classification results of MELD-DA, MIntRec, and MIntRec 2.0. To further analyze the performance of our method, we conduct a fine-grained analysis on intent taxonomies, including hard and non-hard ones. As shown in Table 4, we select 4 well-performed non-hard taxonomies: 'Thank', 'Apologize', 'Agree' and 'Greet'. DuoDN achieves the best performance over all baselines, which indicates its capability of precisely grasping the important semantic information from multimodal inputs. More importantly, when comparing

Methods	Non-hard				Hard			
	Thank	Apologize	Agree	Greet	Taunt	Flaunt	Oppose	Joke
MAG-BERT♣	96.52	97.76	91.60	91.06	15.78	47.09	33.97	37.54
MulT♣	96.83	97.93	92.23	86.65	26.12	48.91	34.68	33.95
MISA♣	98.03	97.78	92.05	82.71	22.15	46.44	36.15	38.74
SDIF-DA*	97.96	98.11	92.31	86.96	25.00	44.44	30.00	55.56
TCL-MAP	97.00	97.70	93.10	90.10	17.20	50.80	35.90	29.00
DuoDN (ours)	**99.06**	**98.63**	**94.42**	**91.11**	**32.26**	**56.62**	**41.38**	**58.38**
Human♣	96.90	96.15	87.21	94.15	65.55	78.10	69.04	72.22

Table 4: F1-score comparison between baselines and DuoDN for intent taxonomies on MIntRec. ♣ denotes the results from Zhou et al. (2023). * are from our re-implementation.



(a) Visualization on MIntRec

(b) Visualization on MELD-DA

Modality-oriented: ● $\boldsymbol{h}_T$ ● $\boldsymbol{h}_V$ ● $\boldsymbol{h}_A$
Semantics-oriented: ● $\boldsymbol{h}_{V,tv}$ ● $\boldsymbol{h}_{T,tv}$ ● $\boldsymbol{h}_{A,ta}$ ● $\boldsymbol{h}_{T,ta}$

Figure 3: Visualization of the semantics-oriented and modality-oriented subspaces in the testing set of MIntRec and MELD-DA datasets using UMAP projections. The left subgraph of each group shows the raw data distribution and the right subgraph displays the data distribution after training. Observations on MIntRec 2.0 are similar.

with human performance, our method still maintains a better performance except for ‘Greet’. Nevertheless, for the hard ones like ‘Taunt’, ‘Flaunt’, ‘Oppose’ and ‘Joke’, although we still keep the best position on the table, the results are far less satisfying. The model struggles with intents that involve complex human emotions and social cues, such as ’Taunt’ and ’Joke’. On one hand, these intents require an understanding of nuance and context that may not be fully captured by the current model architecture or the dataset used for training. On the other hand, the dataset bias within MELD-DA could be influencing the model’s ability to learn certain intents, which is a common issue in machine learning and affects the generalizability and fairness of the model.

4.6 Visualization Analysis of Disentangled Representation Learning

Understanding the subspace distributions of the semantics- and modality-oriented representations is essential to further improving the overall performance of the proposed model. To this end, we visualize the dual-oriented subspaces on MIntRec and MELD-DA using UMAP (McInnes et al., 2018) projection. Before applying the semantic-level contrastive learning and counterfactual intervention method, the distributions of $\boldsymbol{H}_{sem}$ barely overlap. The commonness between modalities is not properly learned. However, after training, there is a clustering phenomenon observed in the distributions of $\boldsymbol{H}_{T,tv}$ and $\boldsymbol{H}_{V,tv}$, as well as $\boldsymbol{H}_{T,ta}$ and $\boldsymbol{H}_{A,ta}$. Furthermore, the semantics-oriented representations share a small but significant subspace, which contributes to our superior performance. This indicates that our proposed SL and CIM effectively align the distributions of different modalities and minimize the gap among them. In addition, each modality-oriented subspace is also separable, optimized by $\mathcal{L}_{mod}^{IE}$. The above observations prove that our method captures the semantic commonality and modality heterogeneity of different modalities.

5 Conclusion

In this paper, we propose a novel Dual-oriented Disentangled Network with Counterfactual Intervention (DuoDN) for multimodal intent detection. Through the dual-oriented disentanglement, DuoDN creates multimodal semantic spaces and optimizes the semantics-oriented representations with semantic-level contrastive learning. Additionally, we introduce a counterfactual intervention module as an additional supervision to highlight the semantics-oriented and modality-oriented representation, leading to a better modeling of the task. Extensive experiments on three benchmark datasets demonstrate DuoDN's effectiveness.

Limitations

Despite the notable superiority of our proposed method over existing SOTA approaches, it is imperative to acknowledge and address several challenges in future research endeavors. Firstly, the categories of the three datasets are unevenly distributed. Since all three datasets suffer from a dataset bias problem, we plan to discover more approaches for data augmentation and dataset debiasing in the future. Secondly, our DuoDN is only suited for single-turn conversations, suffering from difficulties in handling multi-turn multi-party discussions. Although multi-turn conversations are more relevant to real life, effectively leveraging context information remains a substantial challenge. These limitations present a critical area for further investigation in our subsequent research efforts.

Acknowledgement

We would like to thank all reviewers for their insightful comments. This paper was partially supported by NSFC (No: 62176008). Special acknowledgements are given to AOTO-PKUSZ Joint Research Center for its support.

References

Bhuvan Agrawal, Markus Müller, Samridhi Choudhary, Martin Radfar, Athanasios Mouchtaris, Ross McGowan, Nathan Susanj, and Siegfried Kunzmann. 2022. Tie your embeddings down: Cross-modal latent spaces for end-to-end spoken language understanding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7157–7161. IEEE.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. *arXiv preprint arXiv:1907.05789*.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in neural information processing systems*, 29.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, Rui Wang, Feng Xia, and Wenwu Zhu. 2021. Curriculum disentangled recommendation with noisy multi-feedback. *Advances in Neural Information Processing Systems*, 34:26924–26936.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Victor Machado Gonzaga, Nils Murrugarra-Llerena, and Ricardo Marcacini. 2021. Multimodal intent classification with incomplete modalities using text embedding propagation. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 217–220.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR.

Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. 2022. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3394–3402.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

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

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. 2023. Sdif-da: A shallow-to-deep interaction framework with data augmentation for multi-modal intent detection. *arXiv preprint arXiv:2401.00424*.

Yinghui Huang, Hong-Kwang Kuo, Samuel Thomas, Zvi Kons, Kartik Audhkhasi, Brian Kingsbury, Ron Hoory, and Michael Picheny. 2020. Leveraging unpaired text data for training end-to-end speech-to-intent systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7984–7988. IEEE.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Guangyao Li, Wenxuan Hou, and Di Hu. 2023. Progressive spatio-temporal perception for audio-visual question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7808–7816.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. 2020. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869.

Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Mengshi Qi, Jie Qin, Yi Yang, Yunhong Wang, and Jiebo Luo. 2021. Semantics-aware spatial-temporal binaries for cross-modal video retrieval. *IEEE Transactions on Image Processing*, 30:2989–3004.

Mengshi Qi, Jie Qin, Xiantong Zhen, Di Huang, Yi Yang, and Jiebo Luo. 2020. Few-shot ensemble learning for video classification with slowfast memory networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3007–3015.

Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. 2019. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2617–2633.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. *arXiv preprint arXiv:1909.02188*.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A survey on spoken language understanding: Recent advances and new frontiers. *arXiv preprint arXiv:2103.03095*.

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

Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1025–1034.

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*, pages 4361–4372.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

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

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

Pei Wang and Nuno Vasconcelos. 2020. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8990.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Shaoxaing 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. *arXiv preprint arXiv:2305.14652*.

Yuxin Xie, Zhihong Zhu, Xianwei Zhuang, Liming Liang, Zhichang Wang, and Yuexian Zou. 2024. Gpa: Global and prototype alignment for audio-text retrieval. In *Interspeech 2024*, pages 5078–5082.

Bowen Xing and Ivor W Tsang. 2022. Co-guiding net: Achieving mutual guidances between multiple intent detection and slot filling via heterogeneous semantics-label graphs. *arXiv preprint arXiv:2210.10375*.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*.

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

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

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

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

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. 2021. Textoir: An integrated and visualized platform for text open intent recognition. *arXiv preprint arXiv:2110.15063*.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Wenrui Li, Yanting Chen, et al. 2024. Mintrec2. 0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. *arXiv preprint arXiv:2403.10943*.

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

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. 2023. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. *arXiv preprint arXiv:2312.14667*.

Zhihong Zhu, Weiyuan Xu, Xuxin Cheng, Tengtao Song, and Yuexian Zou. 2023. A dynamic graph interactive framework with label-semantic injection for spoken language understanding. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A More Details about Benchmark Datasets

MIntRec is a detailed dataset for detecting multimodal intents, with 2,224 high-quality samples in

20 categories. It's divided into three parts: 1,334 training samples, 445 validation samples, and 445 testing samples.

MELD-DA is a large-scale dataset for classifying dialogue actions, containing 9,988 samples in 12 common dialogue act categories. It's split into 6,991 samples for training, 999 for validation, and 1,998 for testing. Each sample is labeled with one of 12 common dialogue act categories.

MIntRec 2.0 is a benchmark dataset for multimodal intent recognition in multi-party conversations. It consists of 1,245 high-quality dialogues with 15,040 samples, each annotated with one of 30 fine-grained classes that include text, video, and audio modalities. It contains over 9,300 in-scope samples and more than 5,700 out-of-scope samples appearing in multi-turn contexts, which occur naturally in real-world open scenarios, making it more practical. The dataset is divided into training, validation, and testing sets at an approximate ratio of 7:1:1 for both utterances and dialogues. In this paper, we consider all dialogue as a single-turn dialogue for further benchmarking and analysis.

MOSI is a collection of YouTube monologues consisting of 2,199 movie samples, which are separated into 1,284 training samples, 229 validation samples, and 686 testing samples.

MOSEI is an improvement over MOSI with a total of 23,453 video clips, spanning 1,000 distance speakers.

B Structual Causal Model Formulation

To represent the causality links among input $\boldsymbol{X}$, intermediate hidden state $\boldsymbol{H}$, and prediction $\boldsymbol{Y}$, we formulate them with the Structural Causal Model (SCM) (Pearl et al., 2016; Pawlowski et al., 2020) $\mathcal{G} = \{\boldsymbol{V}, \boldsymbol{E}\}$, where $\boldsymbol{N}$ and $\boldsymbol{E}$ represent the set of variable nodes and causal correlations, respectively. The causality links denotes: *cause→ effect*. For instance, the causality could be formulated as:

- $\boldsymbol{X} \rightarrow \boldsymbol{Y}$: the conventional model.
- $\boldsymbol{X} \rightarrow \boldsymbol{H}$: the model produces the corresponding attention.
- $\boldsymbol{X} \rightarrow \boldsymbol{Y} \leftarrow \boldsymbol{H}$: the final prediction $\boldsymbol{Y}$ is determined by $(\boldsymbol{X}, \boldsymbol{H})$ jointly.

As demonstrated in the main body, we consider the message-passing process $\boldsymbol{X} \rightarrow \boldsymbol{H}_{sem} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{X} \rightarrow \boldsymbol{H}_{mod} \rightarrow \boldsymbol{Y}$ as SCM, where we intervene by substituting the input $\boldsymbol{X}$ with confounder $\boldsymbol{X}^*$. With SCM, the causality links between the variables can be directly analyzed via variable intervention, which means manipulating the value of specific variables and then observing the effect.

C Counterfactual Intervention Formulation

In traditional causal models' training, the variable $\boldsymbol{H}$ predicts the output $\hat{Y}$ by only sensing the effective properties of the input $\boldsymbol{X}$. However, this message-passing process often fails to receive enough attention during the end-to-end training process, leading to sub-optimal performance of the network. To tackle this issue, we propose leveraging the counterfactual intervention $Do(\cdot)$, which can highlight the causal link between the confounders and the factual hidden state.

Counterfactual intervention $Do(\cdot)$ is a method used to examine the impact of specific variables. The term counterfactual means 'counter to the facts', and it involves an imaginary intervention that replaces the variables' state, which is not possible to occur in the real world. For instance, $Do(\boldsymbol{H} = \boldsymbol{C})$ implies that the counterfactual variable $\boldsymbol{C}$ is assigned to $\boldsymbol{H}$, which breaks the causality link between $\boldsymbol{H}$ and all of its parent nodes. This action forces the variable to be affected by the confounder $\boldsymbol{C}$. As a result, the direct causality link between the factual state $\boldsymbol{H}$ and the prediction $\boldsymbol{Y}$ can be examined. Specifically, the process $\mathcal{H}(\cdot)$ produces the value of the factual state $\boldsymbol{H}$, while the process $\mathcal{C}(\cdot)$ produces the value of the counterfactual intervention $\boldsymbol{C}$.

$$\boldsymbol{H} = \mathcal{H}(\boldsymbol{X}) = \{\boldsymbol{H}_1, \boldsymbol{H}_2, ..., \boldsymbol{H}_n\}, \tag{21}$$

$$\boldsymbol{C} = \mathcal{C}(\boldsymbol{X}) = \{\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, ..., \boldsymbol{X}_n^*\}, \tag{22}$$

where n denotes the channel number of $\boldsymbol{H}$ and $\boldsymbol{C}$ to control the capacity to perceive the sample-wise properties. Using this, we can analyze the direct causality link between $\boldsymbol{H}$ and $\boldsymbol{Y}$ while excluding the confounders, through the likelihood of counterfactual intervention denoted as $\boldsymbol{P}(\boldsymbol{Y}|Do(\boldsymbol{H} = \boldsymbol{C}))$. We can formulate the likelihood of factual attention as $\boldsymbol{Y}_f$ and counterfactual intervention as $\boldsymbol{Y}_{cf}$, both of which are calculated as follows,

$$\boldsymbol{Y}_f = \boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{H}), \tag{23}$$

$$\boldsymbol{Y}_{cf} = \boldsymbol{P}(\boldsymbol{Y}|Do(\boldsymbol{H} = \boldsymbol{C})) = \boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{H}^*). \tag{24}$$

The former variable $\boldsymbol{Y}_f$ plays a crucial role in creating a model that can generate distinct and understandable representations, while the latter variable

Methods	MOSI					MOSEI				
	MAE	Corr	Acc-7	Acc-2	F1	MAE	Corr	Acc-7	Acc-2	F1
LMF	0.917	0.695	33.20	-/82.50	-/82.40	0.623	0.677	48.00	-/82.00	-/82.10
MFM	0.877	0.706	35.40	-/81.70	-/81.60	0.568	0.717	51.30	-/84.40	-/84.30
MulT	0.861	0.711	40.00	81.50/84.10	80.60/83.90	0.580	0.703	51.80	-/82.50	-/82.30
MISA	0.804	0.764	42.30	80.79/82.10	80.77/82.03	0.568	0.724	52.20	82.59/84.23	82.67/83.97
Self-MM	0.712	0.795	45.79	82.54/84.77	82.68/84.91	0.529	0.767	53.46	82.68/84.96	82.95/84.93
MMIM	0.700	0.800	**46.65**	84.14/86.06	84.00/85.98	0.526	0.772	54.24	82.24/85.97	82.66/85.94
FDMER	0.724	0.788	44.10	-/84.60	-/84.70	0.536	**0.773**	54.10	-/86.19	-/85.80
DBF	**0.693**	**0.801**	44.80	**85.10/86.90**	**85.10/86.90**	**0.523**	0.772	**54.20**	**84.30/86.40**	**84.80/86.20**
DuoDN (ours)	0.700	0.795	45.88	84.68/86.01	84.79/85.04	0.528	0.768	54.00	84.02/86.11	84.36/85.83

Table 5: Performance on MOSI and MOSEI for multimodal sentiment analysis. *Acc-2* denotes the accuracy over negative/non-negative, and *F1* corresponds to negative/positive.

$\boldsymbol{Y}_{cf}$ represents the context-specific confounders that should be eliminated from the prediction. We then compute the difference in likelihood between the factual and counterfactual states to determine the indirect causality effect $\boldsymbol{Y}_{ie}$ between the factual inputs $\boldsymbol{X}$ and their corresponding prediction $\boldsymbol{Y}$.

$$\boldsymbol{Y}_{ie} = \mathbb{E}(\boldsymbol{Y}_f) - \mathbb{E}(\boldsymbol{Y}_{cf}). \tag{25}$$

Maximizing the likelihood difference $\boldsymbol{Y}_{ie}$ could force the network to focus on factual state learning instead of collapsing into sub-optimal performance. Thus, counterfactuals can be regarded as additional supervision to illuminate the significance of specific reasoning features.

D Generalization Performance on Multimodal Sentiment Analysis

To verify the generalization performance of DuoDN, we conduct experiments on the MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018) benchmark datasets for multimodal sentiment analysis. We compared our model with several advanced methods, including LMF, MFM (Tsai et al., 2018), MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), FDMER (Yang et al., 2022), and DBF (Wu et al., 2023).

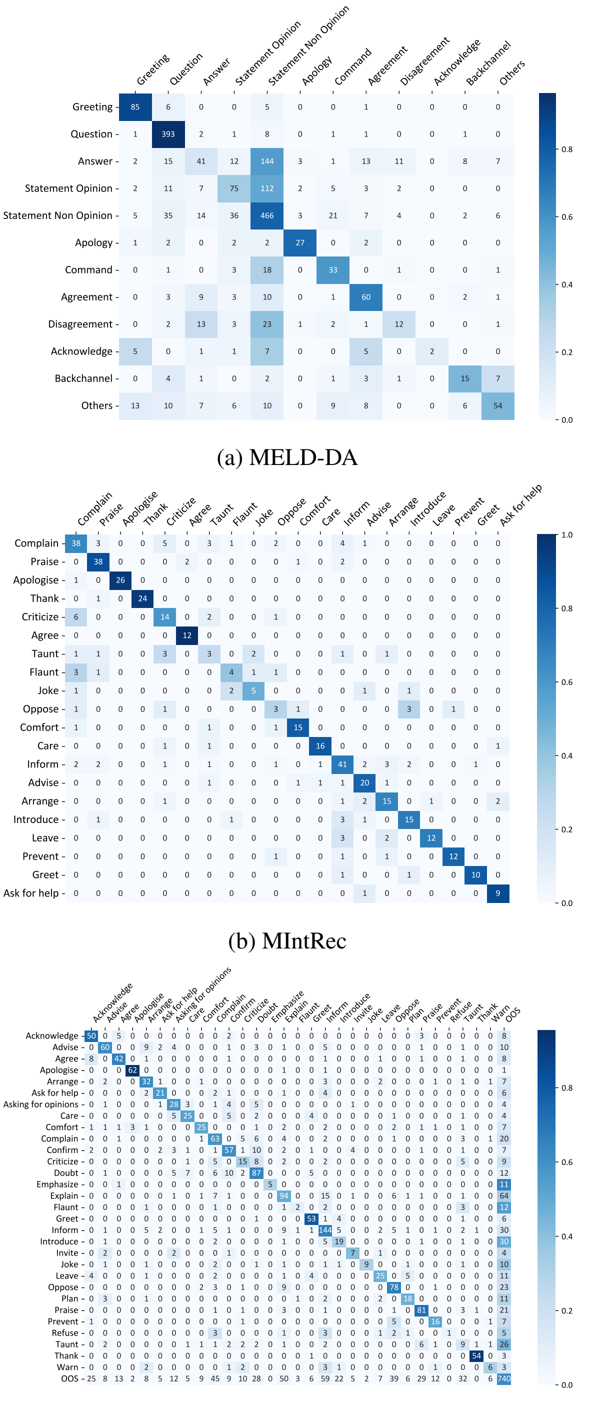
The results, presented in Table 5, show that our model performs competitively against all baselines. Notably, our model achieved the second-best results for the metrics of *Acc-2* and *F1* on both datasets, demonstrating excellent generalization ability. Our approach, which utilizes modality- and semantic-specific information and incorporates causal inference, proves effective and generalizable across different multimodal tasks. DuoDN also excels in fine-grained analysis, as reflected in its superior *Acc-7* performance. Overall, this study provides strong evidence of DuoDN's consistent ability to enhance performance in multimodal sentiment analysis.

E Error Analysis

Results Visualization on MELD-DA On one hand, DuoDN achieves a satisfying performance on the classification of '`Statement Non Opinion`' and '`Question`', which is attributed to their clear semantics. On the other hand, when it comes to ambiguous actions such as '`Answer`', '`Acknowledge`' and '`Others`', our method appears to offer limited assistance in identifying the actions. The accuracies are only 15.95%, 9.52% and 21.95%, respectively. From another perspective, the presence of dataset bias in MELD-DA may also affect the precise learning of complex taxonomies.

Results Visualization on MIntRec Along the diagonal, the intentions '`Complain`', '`Praise`', '`Apologise`', '`Thank`', and '`Inform`' have the highest correct predictions of 38, 38, 26, 24, and 41 respectively. While this shows that the model is better at identifying these intentions, it also shows that these intents are more prevalent in MIntRec. Some classes such as '`Ask for help`', '`Greet`', '`Prevent`', '`Leave`', '`Introduce`', and '`Arrange`' have very few instances. To further improve the results, we need to understand the context in which the data was collected and how each action is defined. For example, '`Joke`' being misclassified as '`Taunt`' could be due to a thin line between them in the way they are used in the dataset.

Results Visualization on MIntRec 2.0 To better learn out-of-scope detection in open-world scenarios, we visualize the intent classification trained with out-of-scope samples. As we can see in Figure 4c, most are correctly classified, except for a few scattered along the diagonal. More importantly, most of the negative results are misclassified as out-of-scope, which contains the most samples. The results indicate that we need to pay more attention to the out-of-scope utterances, which commonly occur in dialogue systems and are crucial for improving system robustness.



(a) MELD-DA

(b) MIntRec

(c) MIntRec 2.0

Figure 4: Visualization of intents classification in the testing set of MELD-DA, MIntRec, and MIntRec 2.0. The value on i-th row and j-th column entry indicates the number of samples with the true label being i-th intent and the predicted label being j-th intent, while the color displays the percentage.