

TECO: Improving Multimodal Intent Recognition with Text Enhancement through Commonsense Knowledge Extraction

Quynh-Mai Thi Nguyen, Lan-Nhi Thi Nguyen, Cam-Van Thi Nguyen*

Faculty of Information Technology

VNU University of Engineering and Technology

{21020125, 21020372, vanntc}@vnu.edu.vn

Abstract

The objective of multimodal intent recognition (MIR) is to leverage various modalities—such as text, video, and audio—to detect user intentions, which is crucial for understanding human language and context in dialogue systems. Despite advances in this field, two main challenges persist: (1) *effectively extracting and utilizing semantic information from robust textual features*; (2) *aligning and fusing non-verbal modalities with verbal ones effectively*. This paper proposes a **Text Enhancement with Commonsense Knowledge Extractor (TECO)** to address these challenges. We begin by extracting relations from both generated and retrieved knowledge to enrich the contextual information in the text modality. Subsequently, we align and integrate visual and acoustic representations with these enhanced text features to form a cohesive multimodal representation. Our experimental results show substantial improvements over existing baseline methods.

1 Introduction

Intent recognition plays a vital role in natural language understanding. While prior attempts focused on a single modality, e.g., text, for extraction (Hu et al., 2021), real-world scenarios involve intricate human intentions that require the integration of information from speech, tone, expression, and action. Recently, multimodal intent recognition (MIR) performed computationally is a very interesting and challenging task to be explored. To effectively leverage the information from various modalities, numerous methods have been proposed for MIR. As an alternative, (Tsai et al., 2019); (Rahman et al., 2020) proposed frameworks using transformer-based techniques to integrate information from different modalities into a unified feature.

*Corresponding author. Cam-Van Thi Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.TS147.

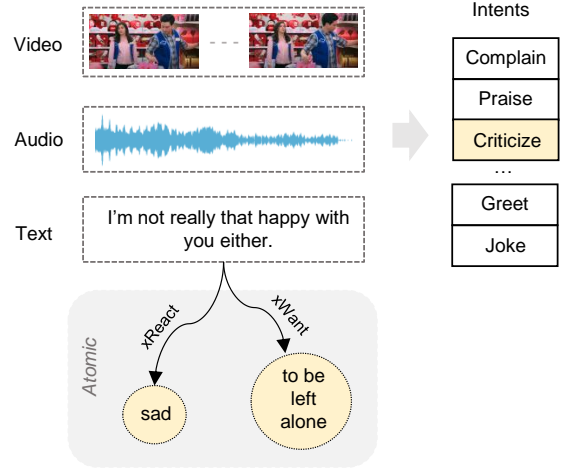


Figure 1: An example of integrating commonsense knowledge for multi-intent recognition provides awareness about implicit context which relates to the utterance’s intention.

Moreover, (Zhou et al., 2024) introduced a token-level contrastive learning method with a modality-aware prompting module; (Huang et al., 2024) proposed a shallow-to-deep transformer-based framework with ChatGPT-based data augmentation strategy, achieving an impressive result. Despite the advances, we suppose that existing MIR models still suffer from several challenges: (1) how to explore the semantic information from the contextual features effectively; (2) the limitation in aligning and fusing features of different modalities.

To address the above challenges, we introduce a framework called Text Enhancement with Commonsense Knowledge Extractor (**TECO**). Our model comprises three main components: a Commonsense Knowledge Extractor (COKE), a Textual Enhancement Module (TEM), and a Multimodal Alignment Fusion (MAF). Our main idea is to explore rich and comprehensive contextual features and then incorporate them with non-verbal features (image, audio) to predict the reasonable utterance

of the participants. COKE combines both retrieved and generated commonsense knowledge to capture relational features, whereas TEM utilizes a dual perspective learning module and a textual enhancing fusion to integrate them into the text feature. Finally, we adopt MAF to effectively fuse features from three modalities into multimodal knowledge-enhanced representations of utterances.

Our contributions are summarized as follows:

- We propose the TECO model, featuring a Text Enhancement Module (TEM) with commonsense knowledge extraction to effectively leverage semantic information from textual input.
- TECO incorporates Dual Perspective Learning to integrate and harmonize relation perspectives and aligns non-verbal modalities with verbal ones for consistent multimodal representation.
- Experimental results and detailed analyses on the challenging MIntRec dataset demonstrates the superior performance of our TECO model in multimodal intent detection.

2 Related Works

2.1 Commonsense Knowledge

Commonsense reasoning utilizes the basic knowledge that reflects our natural understanding of the world and human behavior, which is crucial for interpreting the latent variables of a conversation. Recently, COMET (Bosselut et al., 2019) has achieved impressive performance when investigating and transferring implicit knowledge from a deep pre-trained language model to generate explicit knowledge in commonsense knowledge graphs. The seminal works utilize COMET to guide the participants through their reasoning about the content of the conversation, dialog planning, making decisions, and many reasoning tasks. SHARK (Wang et al., 2023) uses a pre-trained neural knowledge model COMET-ATOMIC (Hwang et al., 2021) to extract emotion utterance by generating novel commonsense knowledge tuples, CSDGCN (Yu et al., 2023) proposed using COMET to clearly depict how external commonsense knowledge expressions within the context contributes to sarcasm detection, R^3 (Chakrabarty et al., 2020) retrieve relevant context for the sarcastic messages based on commonsense knowledge.

Sentence-BERT (Reimers and Gurevych, 2019) uses siamese and triplet network structure to capture semantically meaningful sentence features that can be compared using cosine-similarity. In this paper, we incorporate two views from generative and retrieved relations to enrich context information via two pre-trained models, COMET and SBERT.

2.2 Multimodal Fusion

Multimodal Fusion is an active area of research with various proposed methods. Prior studies based on transformer, MULT (Tsai et al., 2019) directly attend to elements in other modalities and capture long-range crossmodal events. However, it does not handle modality non-alignment by simply aligning them. Moreover, MAG-BERT (Rahman et al., 2020) proposed an efficient framework for fine-tuning BERT (Devlin, 2018) and XLNet (Yang, 2019) for multimodal input and MISA (Hazari et al., 2020) projects each modality to two distinct subspaces, which provide a holistic view of the multimodal data. To effectively fuse different modalities’s features and alleviate the data scarcity problem, SDIF-DA (Huang et al., 2024) introduced a shallow-to-deep interaction framework using a hierarchical and a transformer module. Recent researches attempt to extract more information from textual input, Prompt Me Up (Hu et al., 2023) proposed innovative pre-training objects for entity-object and relation-image alignment, extracting objects from images and aligning them with entity and relation prompts. To leverage the limitations in learning semantic features, TCL-MAP (Zhou et al., 2024) develops a token-level contrastive learning method with a modality-aware prompting module.

3 Methodology

3.1 Problem Statement and Model Overview

Problem Statement. Multi-modal intent recognition aims to analyze various modalities such as expression, body movement, and tone of speech to understand a user’s intent. Given an input text $T = \{t_1, t_2, \dots, t_{l^S}\}$ with the corresponding image V and audio A , where l^S is the length of the text sequence, our model is supposed to classify given text into correct intent category $i \in \mathbb{I} = \{i_1, i_2, \dots, i_N\}$. The set \mathbb{I} contains the pre-defined intent types, and N represents the number of utterances.

Model Overview. Figure 2 describes the architecture of our model, which comprises three components. The input sentence is converted into

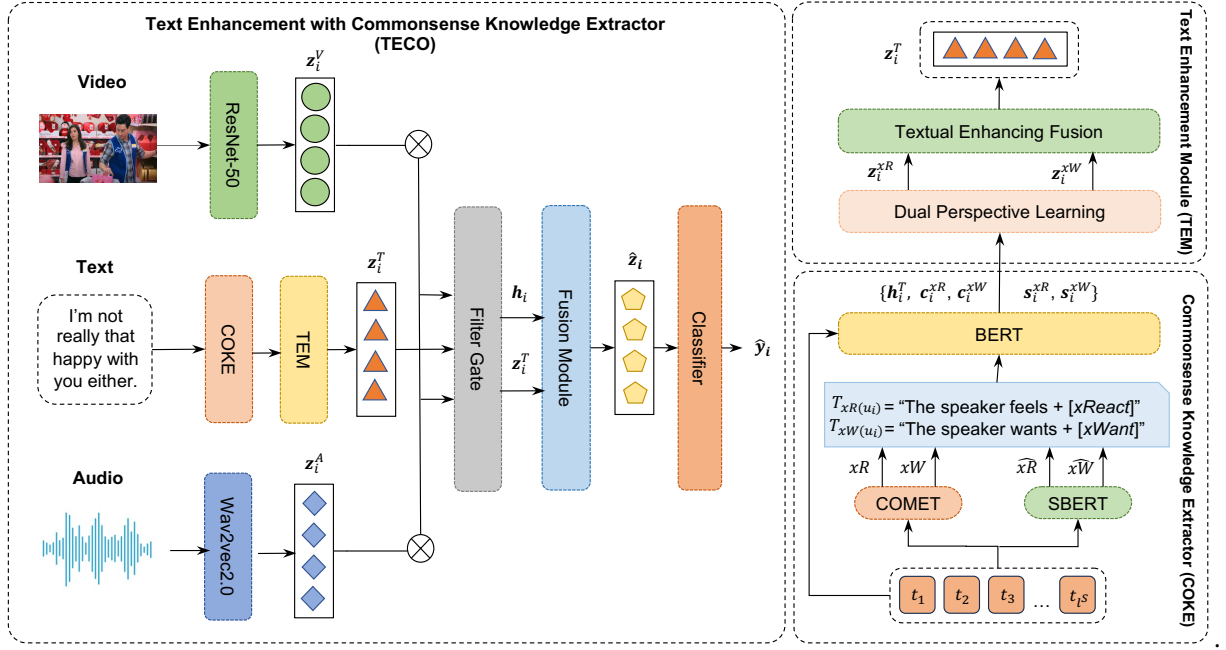


Figure 2: Overall architecture of our model is illustrated in the left part. The lower right part describes the flow of the Commonsense Knowledge Extractor (COKE), and the upper one shows details of the Text Enhancement Module (TEM), which integrates relation features into textual representations using commonsense knowledge and a dual perspective learning module.

vector representations using an encoding context module. Next, in the Textual Enhancement Module (TEM), we utilize a commonsense reasoning module to extract relevant knowledge and convert it into vector representations. Subsequently, the output vector is put into a dual mechanism to obtain a single representation.

We also extract features from audio segments and video segments by using encoder mechanisms. After each extracted feature is aligned with the textual information, we concatenate the textual feature with the visual and acoustic information and utilize them to compute two filter gates, which emphasize relevant information from visual and acoustic modalities based on the textual input. We then separately feed each obtained feature into a fusion module. Finally, in the prediction stage, we perform a classifier operation to get the final utterance detection result.

3.2 Feature Encoders

For each utterance u_i , we extract multimodal features from three different modalities: text, vision, and audio.

Textual Encoder. The pre-trained BERT language model (Devlin, 2018) which achieves excellent performance in Natural Language Processing (NLP) is applied to extract text features. For each

input sentence t_i , we obtain the token embeddings from the last hidden layer of the BERT Encoder:

$$\mathbf{h}_i^T = \text{TextEncoder}(t_i) \quad (1)$$

where TextEncoder is BERT Encoder, $\mathbf{h}_i^T \in \mathbb{R}^{l^S \times d}$ refers to the text embedding of text sentence t_i , l^S is the length of text sentence, and d denotes the feature dimension.

Visual Encoder. We follow the approach used in previous work (Zhang et al., 2022) to process video segments. By leveraging a pre-trained Faster R-CNN (Ren et al., 2015) with the backbone ResNet-50 (Koonce and Koonce, 2021), the vision feature embeddings are extracted as follows:

$$\mathbf{h}_i^V = \text{VisualEncoder}(v_i) \quad (2)$$

where VisualEncoder is Faster R-CNN, $\mathbf{h}_i^V \in \mathbb{R}^{l^V \times d^V}$ denotes the vision embedding of video segment v_i , l^V is the length of video segment, and d^V refers to the vision feature dimension.

Acoustic Encoder. To extract the acoustic embeddings, we utilize a pre-trained model wav2vec 2.0 (Schneider et al., 2019), which employs self-supervised learning to generate strong representations for speech recognition. The formula is shown as follows:

$$\mathbf{h}_i^A = \text{AcousticEncoder}(a_i) \quad (3)$$

where AcousticEncoder refers to wav2vec 2.0, $\mathbf{h}_i^A \in \mathbb{R}^{l^A \times d^A}$ denotes the acoustic embedding of audio segment a_i , l^A is the audio segment’s length, and d^A denotes the acoustic feature dimension.

3.3 Commonsense Knowledge Extractor (COKE)

For each utterance, we utilize a commonsense knowledge graph combined with two pre-trained models to obtain relational features. Subsequently, integrating them into textual features to enhance textual information.

Relation Generation. We put each utterance through a pre-trained generative model COMET¹ (Bosselut et al., 2019), which is able to produce rich and diverse commonsense knowledge relying on a seed set of knowledge tuples. A knowledge base ATOMIC² (Hwang et al., 2021) is used as a knowledge seed set to generate phrases of several relation types. Among nine relation types, we choose $xReact$ and $xWant$ as generative relation representations. For example, given the input utterance “I’m not really that happy with you either” and get the output $xReact$ and $xWant$ are “sad” and “to be left alone”, respectively.

Relation Retrieval. To retrieve relational knowledge, we apply SBERT (Reimers and Gurevych, 2019) to compute the similar score between each utterance and each sentence in the ATOMIC dataset. After that, we select the phrases under the two relation types $xReact$ and $xWant$ of the most similar sentence as retrieved relation representations. In particular, the $xReact$ and $xWant$ phrases of the utterance “wait, it’s- hey, stop... stop!” are “frustrated” and “to scold someone”, respectively.

Relation Encoding. After obtaining the relation phrases, we put them into a combined template in order to receive the complete sentence S_{rel} . The combined template is formalized as:

$$\begin{aligned} T_{xR}(u_i) &= \text{“The speaker feels } [xReact].\text{”} \\ T_{xW}(u_i) &= \text{“The speaker wants } [xWant].\text{”} \end{aligned} \quad (4)$$

where $T(u_i)$ refers to the combined template of each relation type corresponding to the utterance u_i .

The complete sentences of generative and retrieved relation are separately fed to the BERT encoder to gain relation features. Finally, for each ut-

terance u_i , we obtain four relation representations including $\mathbf{c}_i^{xR}, \mathbf{c}_i^{xW}, \mathbf{s}_i^{xR}, \mathbf{s}_i^{xW} \in \mathbb{R}^{l^R \times d}$, where l^R denotes the length of the complete relation sentence.

3.4 Textual Enhancement Module

To take advantage of commonsense knowledge, we employ a Textual Enhancement Module (TEM) which integrates the relation features into textual features to enrich textual representations.

Dual Perspective Learning. We apply a dual perspective learning mechanism to perform relation fusion from two different views: generative and retrieved knowledge. First, we calculate learnable weight through a linear layer for each relation type. The formula is defined as follows:

$$\begin{aligned} \alpha_i^{xR} &= \text{SoftMax}(f_L([\mathbf{h}_i^T, \mathbf{c}_i^{xR}, \mathbf{s}_i^{xR}])) \\ \alpha_i^{xW} &= \text{SoftMax}(f_L([\mathbf{h}_i^T, \mathbf{c}_i^{xW}, \mathbf{s}_i^{xW}])) \end{aligned} \quad (5)$$

where $\alpha_i^{xR}, \alpha_i^{xW}$ is the learnable weight corresponding to $xReact$ and $xWant$ relation, and f_L denotes the linear layer.

Next, the relation fusion features are computed as follows:

$$\begin{aligned} \mathbf{h}_i^{xR} &= \alpha_i^{xR} \cdot \mathbf{c}_i^{xR} + (1 - \alpha_i^{xR}) \cdot \mathbf{s}_i^{xR} \\ \mathbf{h}_i^{xW} &= \alpha_i^{xW} \cdot \mathbf{c}_i^{xW} + (1 - \alpha_i^{xW}) \cdot \mathbf{s}_i^{xW} \end{aligned} \quad (6)$$

where $\mathbf{h}_i^{xR}, \mathbf{h}_i^{xW} \in \mathbb{R}^{l^R \times d}$.

Textual Enhancing Fusion. After obtaining the relation fusion features, we integrate them into the text feature by learning a trainable weight and tuning a hyper-parameter fused relation. For details, the formula is described as follows:

$$\begin{aligned} \mathbf{z}_i^{xR} &= \mathbf{h}_i^T + \mathbb{W} \mathbf{h}_i^{xR} \\ \mathbf{z}_i^{xW} &= \mathbf{h}_i^T + \mathbb{W} \mathbf{h}_i^{xW} \end{aligned} \quad (7)$$

$$\mathbf{z}_i^T = \gamma \cdot \mathbf{z}_i^{xR} + (1 - \gamma) \cdot \mathbf{z}_i^{xW} \quad (8)$$

where $\mathbf{z}_i^T \in \mathbb{R}^{l^S \times d}$ is the text-enhanced feature of utterance u_i , \mathbb{W} denotes the trained weight, and γ refers to the hyper-parameter.

3.5 Multimodal Alignment Fusion

Because of the independent learning of three modalities, we adopt a Multimodal Alignment Fusion (MAF) to align contextual information captured from separated modalities and fuse them to obtain the multimodal knowledge-enhanced representation of utterances.

¹<https://github.com/atcbosselut/comet-commonsense>

²<https://github.com/allenai/comet-atomic-2020/>

First, to align the vision and acoustic feature with the text-enhanced feature, we apply the Connectionist Temporal Classification (CTC) (Graves et al., 2006) module:

$$\mathbf{z}_i^T, \mathbf{z}_i^V, \mathbf{z}_i^A = \text{CTC}(\mathbf{z}_i^T, \mathbf{h}_i^V, \mathbf{h}_i^A) \quad (9)$$

where $\mathbf{z}_i^T \in \mathbb{R}^{l^S \times d}$, $\mathbf{z}_i^V \in \mathbb{R}^{l^V \times d}$, $\mathbf{z}_i^A \in \mathbb{R}^{l^A \times d}$ refer to the aligned features under each modality, and CTC is a module that consists of a LSTM block and a SoftMax function.

Subsequently, we concatenate the text-enhanced feature with visual and acoustic features. These concatenated features are then used to compute two filtering gates, which selectively emphasize relevant information within the visual and acoustic modalities, conditioned by the textual feature. The formulation is as follows:

$$\begin{aligned} \mathbf{g}_i^V &= \text{ReLU}(f_{VT}([\mathbf{z}_i^V \parallel \mathbf{z}_i^T])) \\ \mathbf{g}_i^A &= \text{ReLU}(f_{AT}([\mathbf{z}_i^A \parallel \mathbf{z}_i^T])) \end{aligned} \quad (10)$$

where $\mathbf{g}_i^V, \mathbf{g}_i^A$ are two weighted gates related to the visual and acoustic features, ReLU is an activation function, f_* denotes a linear layer and \parallel is notated for concatenating.

Then, we produce the non-verbal feature by fusing the visual and acoustic features through two gates:

$$\mathbf{h}_i = \mathbf{g}_i^V \cdot f_V(\mathbf{z}_i^V) + \mathbf{g}_i^A \cdot f_A(\mathbf{z}_i^A) \quad (11)$$

where $\mathbf{h}_i \in \mathbb{R}^{l \times d}$, l denotes the length of non-verbal token embeddings and f_* is a linear layer.

Finally, we compute a fused weight β between the text-enhanced feature and the non-verbal feature and then utilize it to create the multimodal feature $\bar{\mathbf{z}} \in \mathbb{R}^{l \times d}$:

$$\beta = \min\left(\frac{\|\mathbf{z}_i^T\|_2}{\|\mathbf{h}_i\|_2} \varepsilon, 1\right) \quad (12)$$

$$\bar{\mathbf{z}}_i = f(\mathbf{z}_i^T + \beta \mathbf{h}_i) \quad (13)$$

where $\|\cdot\|_2$ refers to L_2 normalization, ε is a hyper-parameter, and f denotes a normalized block including a layer normalization and dropout layer.

3.6 Prediction and Loss Function

Prediction. The output of the MAF module $\bar{\mathbf{z}}$ is put through a Classifier to obtain the intent probability distribution. For details, the Classifier contains a pooling layer, a dropout layer, and the last one is a linear layer. The equation is described below:

$$\hat{\mathbf{y}}_i = f_c(\text{Dropout}(\text{Pooler}(\bar{\mathbf{z}}_i))) \quad (14)$$

where $\hat{\mathbf{y}}_i \in \mathbb{R}^N$ denotes the predicted output, N is the number of intent classes, and f_c is a linear layer.

Loss Function. During the training phase, we apply a standard cross-entropy loss to optimize the performance of our model:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\hat{\mathbf{y}}_i)}{\sum_{j=1}^N \exp(\hat{\mathbf{y}}_j)} \quad (15)$$

where B is the batch size, and $\hat{\mathbf{y}}_i$ denotes the predicted label of i^{th} sample.

4 Experiments

4.1 Experimental Settings

Dataset. We conduct experiments on MIntRec (Zhang et al., 2022) dataset which is a fine-grained dataset for multimodal intent recognition. This dataset comprises 2,224 high-quality samples with three modalities: text, vision, and acoustic across twenty intent categories. The dataset is divided into a training set of 1,334 samples, a validation set of 445 samples, and a test set of 445 samples.

Implementation Details. For the implementation of our proposed method, we set the training batch size is 16, while the validation and test batch sizes are both 8. The number of epochs for training is set to 100, and we apply early stopping for 8 epochs. To optimize the parameters, we employ an AdamW (Loshchilov and Hutter, 2017) optimizer with linear warm-up and a weight decay of $1e-2$ for parameter tuning. The initial learning rate is set to $2e-5$ and the hyper-parameter fused relation γ is chosen from $[0.05 : 0.95]$. As sequence lengths of the segments in each modality and relation sentence need to be fixed, we use zero-padding for shorter sequences. l^S, l^V, l^A, l^R are 30, 230, 480, and 30, respectively.

Evaluation Metrics. We use four metrics to evaluate our model performance: accuracy (ACC), F1-score (F1), precision (PREC), and recall (REC). The macro score over all classes for the last three metrics is reported. The higher values indicate improved performance of all metrics.

4.2 Baselines

We compare our framework with several comparative baseline methods:

- **Text Classifier** (Zhang et al., 2022) is a classifier with text-only modality that uses the first special token $[CLS]$ from the last hidden

Table 1: Multimodal intent recognition results on the MIntRec dataset. “Twenty-class” and “Binary-class” denote the multi-class and binary classification. The best performances are highlighted in **bold**, and the underline refers to the second-best ones. Results with * are obtained by reimplemented, while others are taken from the corresponding published paper.

Methods	Twenty-class				Binary-class			
	ACC (%)	F1 (%)	PREC (%)	REC (%)	ACC (%)	F1 (%)	PREC (%)	REC (%)
Text Classifier	70.88	67.40	68.07	67.44	88.09	87.96	87.95	88.09
MAG-BERT	72.65	68.64	69.08	<u>69.28</u>	89.24	89.10	89.10	89.13
MuT	<u>72.52</u>	69.25	70.25	69.24	89.19	89.01	89.02	<u>89.18</u>
MISA	72.29	<u>69.32</u>	70.85	69.24	89.21	89.06	89.12	89.06
SDIF-DA*	71.01	67.77	68.75	67.7	88.76	88.65	88.56	88.77
TCL-MAP*	71.46	68.02	67.84	69.23	<u>89.44</u>	<u>89.26</u>	<u>89.44</u>	89.11
TECO (Ours)	72.36	69.96	<u>70.49</u>	69.92	89.66	89.54	89.5	89.58

layer of the BERT pre-trained model as the sentence representation.

- **MAG-BERT** (Rahman et al., 2020) integrated the two non-verbal features including video and acoustic features into the lexical one by applying a Multimodal Adaptation Gate (MAG) module attached to the BERT structure.
- **MuT** (Tsai et al., 2019) stands for the Multimodal Transformer, an end-to-end model that extends the standard Transformer network (Vaswani, 2017) to learn representations directly from unaligned multimodal streams.
- **MISA** (Hazarika et al., 2020) projected each modality to two distinct subspaces. The first one learns their commonalities and reduces the modality gap, while the other is private to each modality and captures their characteristic features. These representations provide a holistic view of the multimodal data.
- **SDIF-DA** (Huang et al., 2024) is a Shallow-to-Deep Interaction Framework with Data Augmentation that effectively fuses different modalities’ features and alleviates the data scarcity problem by utilizing the shallow interaction and the deep one.
- **TCL-MAP** (Zhou et al., 2024) proposed a modality-aware prompting module (MAP) to align and fuse features from text, video, and audio modalities with the token-level contrastive learning framework (TCL).

4.3 Results

Table 1 describes the results conducted on the intent recognition tasks. Overall, our approach gains significant performances compared to the baselines on the two tasks: binary classification and multi-class classification. Especially, in the binary classification stage, our method outperforms the others across all four metrics. Compared to the second-best methods, the considerable enhancements of 0.25% on accuracy, 0.31% on macro F1-score, 0.67% on precision, and 0.53% on recall indicate the efficiency of our model to leverage multimodal information for understanding real-world context. In the remaining task, our method achieves notable improvements on two metrics macro F1-score and recall, and also gains the second-best result on precision. This observation illustrates the capability of our proposed model in recognizing speakers’ intents within a dialog act.

4.4 Ablation Study

4.4.1 Contribution Analysis of Model Components

To further analyze the contributions of each component to overall performance, we conduct a set of ablation studies including setting model with (1) text and video information (w_{TV}), (2) text and audio features (w_{TA}), and (3) video combined with audio representation (w_{VA}); removing (4) the Text Enhancement Module (w/o_{TEM}), (5) the Multimodal Alignment Fusion module (w/o_{MAF}), and (6) the dual perspective learning by detaching SBERT component (w/o_{dual}).

The important role of the text representation. We explore the role of modalities by removing one modality at a time in ablation studies (1), (2), (3).

Table 2: Ablation experiments of several modules within our model on both multi-class and binary classification stages.

Methods	Twenty-class				Binary-class			
	ACC (%)	F1 (%)	PREC (%)	REC (%)	ACC (%)	F1 (%)	PREC (%)	REC (%)
TECO (Ours)	72.36	69.96	70.49	69.92	89.66	89.54	89.5	89.58
w_{TV}	70.79	66.05	66.35	66.77	88.54	88.35	88.48	88.26
w_{TA}	70.34	66.91	67.49	67.04	88.99	88.85	88.83	88.87
w_{VA}	16.85	3.16	2.46	6.66	52.36	48.28	49.75	49.79
w/o_{TEM}	70.34	64.4	64.43	65.03	88.54	88.45	88.33	88.67
w/o_{MAF}	71.91	68.19	68.67	68.45	87.42	87.33	87.22	87.61
w/o_{dual}	69.44	65.68	66.07	65.83	87.19	87.04	86.99	87.1

As shown in Table 2, the accuracy of our methods decreased seriously when the contextual modality was removed. Particularly, similar drops in performance are not observed then other two modalities are removed, which indicates that textual information has a dominant effect.

The effect of dual perspective learning and textual enhancement module. To explore whether the dual perspective learning, we conduct an experiment (6) that removes retrieved relation from SBERT and remains generative relation extracted from COMET to enhance text representation without dual-view. We can observe that the TECO without dual perspective learning experiences a significant lessening of 4.2% and 2.8% in accuracy for multi-class and binary-class classification, respectively. In addition, we remove features obtained from both COMET and SBERT which is described in experiment (4) to prove the necessary role of commonsense knowledge. We can observe that the final result witnessed a substantial decrease in most metrics indicating that our method is successful in strengthening verbal representation.

MAF works productively in multimodal fusion operation. In experiment (5), we assess the effectiveness of multimodal alignment fusion by discharging both two non-verbal features. As indicated by the results, the performance shows a reduction of more than 2% across most metrics for multi-class. The same trend was witnessed in several metrics for binary classification. The experimental results illustrate that contextual modality plays a critical role in integrating and predicting user’s intents.

4.4.2 Hyper-parameter Analysis

To evaluate the influence of each relation type on our model’s performance, we set up experiments by changing the hyperparameter γ in Equation 8. The

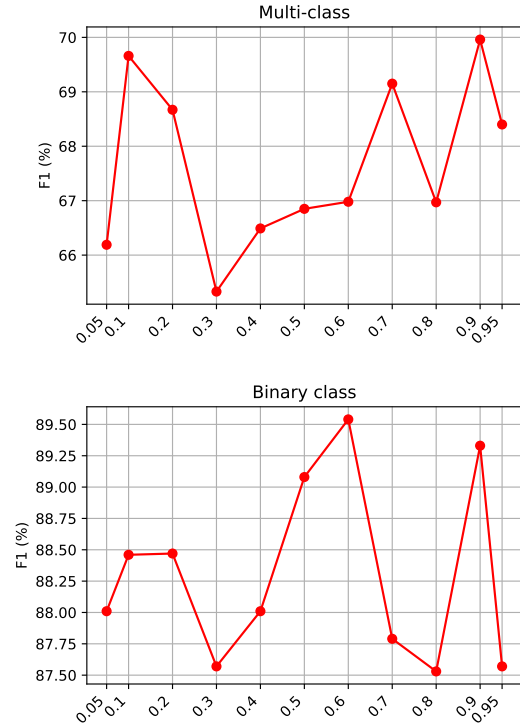


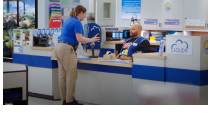





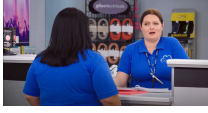

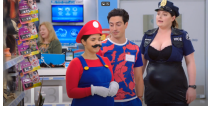

Figure 3: Model performance across different value of γ

results are recorded in Figure 3, which the former is conducted for multi-class classification while the latter is for binary one. We find that macro F1-score is improved at $\gamma = 0.9$ and $\gamma = 0.6$ on multi-class and binary class, respectively. This indicates the relation $xReact$ having more effect on enhancing text representations and boosting the model capability of detecting intention than the relation $xWant$.

4.5 Case Study

To demonstrate the association and impact of the two relations $xReact$ and $xWant$ derived from generative and retrieved knowledge extractor, we

Table 3: The illustration of case studies, where the text with green color indicates the correct prediction, while the other is the incorrect one.

Text	Video	Audio	<i>xReact</i>		<i>xWant</i>		Intent	
			COMET	SBERT	COMET	SBERT	Label	Predicted
"Yeah, those babies look great."			happy	very happy	to have a good time	smile at the baby	Praise	Praise
"And unfortunately, it is supposed to rain."			sad	very worry	to get a umbrella	to stay dry	Complain	Complain
"So thank you all so much for my gifts."			happy	happy	to show appreciation	to accept the givings	Thank	Thank
"Stop, please."			happy	scared	to be a good friend	to get away	Prevent	Oppose
"Hey, we have a problem."			worried	curious	to solve the problem	to make adjustments	Inform	Ask for help

write down several samples in Table 3. The first three examples show the relevance between relation and label intent, which make the donation of producing the correct prediction. Especially, *xReact* tends to express feelings related to intention, while *xWant* is able to generalize the meanings of the sentence. Our COKE module can generate relations more precisely with “expressing emotions” intents such as *Praise*, *Complain*, *Thank* than “achieving goals” such as *Inform*, *Prevent*. In addition, obtaining relations from sentences with clear emotional words is more exact than from those that are brief and ambiguous.

5 Conclusion

In this work, we introduce a Text Enhancement associated with Commonsense Knowledge Extractor (TECO) for multimodal intent recognition. Our model enriches text information by integrating relation information extracted from a commonsense knowledge graph. Thanks to the strength of commonsense knowledge, the implicit contexts of input utterances are explored and utilized to enhance verbal representations. In addition, both visual and acoustic representations are aligned with textual ones to obtain consistent information and then fused together to gain meaningful and rich multimodal features. To evaluate our method’s perfor-

mance, we conducted several experiments and ablation studies on the MIntRec dataset and achieved remarkable results.

References

- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.
- Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R³: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986.
- Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.
- 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.
- Xuming Hu, Junzhe Chen, Aiwei Liu, Shiao Meng, Lijie Wen, and Philip S Yu. 2023. Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5185–5194.
- Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and S Yu Philip. 2021. Semi-supervised relation extraction via incremental meta self-training. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 487–496.
- Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. 2024. Sdif-da: A shallow-to-deep interaction framework with data augmentation for multimodal intent detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10206–10210. IEEE.
- Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6384–6392.
- Brett Koonce and Brett Koonce. 2021. Resnet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- 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.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- 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.
- Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.
- 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.
- A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- Fanfan Wang, Jianfei Yu, and Rui Xia. 2023. Generative emotion cause triplet extraction in conversations with commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3952–3963.
- Z Yang. 2019. Xlnet: Generalized autoregressive pre-training for language understanding. *arXiv preprint arXiv:1906.08237*.
- Zhe Yu, Di Jin, Xiaobao Wang, Yawen Li, Longbiao Wang, and Jianwu Dang. 2023. Commonsense knowledge enhanced sentiment dependency graph for sarcasm detection. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2423–2431.
- 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. 2024. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17114–17122.