LCAN: A Label-Aware Contrastive Attention Network for Multi-Intent Recognition and Slot Filling in Task-Oriented Dialogue Systems

Shuli Zhang[1], Zhiqiang You [1], Xiaoxiang Qi[1], Peng Liu[2], Gaode Wu[3], Kan Xia[3], Shenguang Huang[3],
[1]Hunan University, [2]Guangdong University of Technology,
[3]Ningbo Port Information Communication Co., Ltd.

Abstract

Multi-intent utterances processing remains a persistent challenge due to intricate intent-slot dependencies and semantic ambiguities. Traditional methods struggle to model these complex interactions, particularly when handling overlapping slot structures across multiple intents. This paper introduces a label-aware contrastive attention network (LCAN), a joint modeling approach for multi-intent recognition and slot filling in task-oriented dialogue systems. LCAN addresses this issue by integrating label-aware attention and contrastive learning strategies, improving semantic understanding and generalization in multi-intent scenarios. Extensive experiments on the MixATIS and MixSNIPS datasets demonstrate LCAN's superiority over existing models, achieving improved intent recognition and slot filling performance, particularly in handling overlapping or complex semantic structures in multi-intent settings.

1 Introduction

Task-oriented dialogue (TOD) systems are critical for applications like virtual assistants and customer service automation. A key challenge lies in multi-intent scenarios where single utterances express multiple intents and slots, requiring precise identification for optimal user experience (Wang et al., 2021).

Traditional methods (Xu and Sarikaya, 2013; Zhang and Wang, 2016) treated intent recognition and slot filling as separate tasks. Recent studies reveal their intrinsic connections, driving adoption of joint modeling to enable information sharing, reduce error propagation, and improve efficiency.

Early joint models combined CRFs with CNNs (Xu and Sarikaya, 2013) or GRUs (Zhang and Wang, 2016), later enhanced by attention mechanisms (Liu and Lane, 2016). Pre-trained models revolutionized the field, with Joint BERT (Chen et al., 2019) leveraging BERT embeddings. Subsequent innovations include syntax-enhanced Transformers (Wang et al., 2021) and bilinear attention for parallel intent-slot extraction (Chen et al., 2022a).

Despite advances, multi-intent handling remains challenging. As shown in Figure 1, utterances like "What does ewr stand for, what airlines fly from burbank to denver and airports in New York?" contain multiple intents with distinct slot types. Current models often fail to accurately disentangle and align these overlapping semantic components.

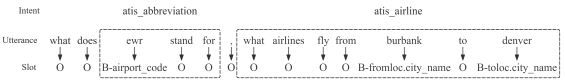

Figure 1: An example of utterance with multiple intents and slots.

As dialogue systems increasingly handle multi-intent scenarios in practical applications, researchers have focused on effective identification and alignment of multiple intents and slots in complex interactions. Kim et al. (2017b) first explored multi-intent recognition in spoken language understanding and introduced joint multi-intent modeling. Gangadharaiah (2019) proposed a joint framework with slot-gating mechanisms, though semantic ambiguity persisted due to single-vector guidance for multiple slot fillings.

To address this limitation, Qin et al. (2020) developed a Graph Attention Network (GAT)-based interaction model establishing fine-grained intent-slot mappings. However, its autoregressive structure constrained bidirectional information capture. Subsequently, Qin et al. (2021a) introduced a global-local graph framework to model complex intent-slot and slot-slot relations, enhancing joint modeling performance.

Xing and Tsang (2022) proposed a heterogeneous graph with mutual "slot-to-intent" and "intent-to-slot" guidance, coupled with label-aware

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25603–25612

decoding for enhanced interaction. Cheng et al. (2023a) designed the SSRN model with explicit intent scope identification and auxiliary tasks to mitigate error propagation.Tu et al. (2023) introduced BiSLU, combining supervised contrastive learning with self-distillation for bidirectional multi-task guidance. Cheng et al. (2023b) further enhanced intent-slot synergy through multi-granular contrastive learning.

Recent advances include Chen et al. (2024)'s two-stage contrastive learning with data augmentation for shared intent relations, and Ma et al. (2024)'s generative task unification using prompt-enhanced fusion. Zhuang et al. (2024) proposed interpretable joint modeling via cross-task information maximization.

Despite progress, significant challenges remain in multi-intent scenarios: (1) Complex intent-slot interactions hinder accurate detection and filling, and (2) Semantic ambiguity persists when intents share slots or exhibit unclear boundaries.

We propose LCAN (Label-Aware Contrastive Attention Network), a novel joint model integrating label-aware attention with contrastive learning to address these issues. LCAN enhances semantic understanding through contrastive intent-slot alignment while employing label-aware attention to reduce input ambiguity. Our framework enables effective information sharing between tasks while minimizing semantic interference.

Experiments on MixATIS and MixSNIPS datasets demonstrate LCAN's superiority over state-of-the-art baselines, particularly in complex multi-intent scenarios. Ablation studies confirm the critical role of contrastive learning, with component-wise analysis revealing performance contributions.

In this paper, we make the following contributions:(1)We propose LCAN, a joint model that combines label-aware attention and contrastive learning to improve multi-intent recognition and slot filling in task-oriented dialogue systems.(2)We demonstrate the effectiveness of LCAN through experiments on two standard benchmark datasets, MixATIS and MixSNIPS, and compare its performance with existing state-of-the-art models.(3)We conduct ablation studies to analyze the contribution of each component of LCAN and show that the contrastive learning modules play a key role in enhancing the model's performance.

2 Related work

2.1 Multi-Intent NLU

In comparison to single-intent detection (Wu et al., 2020), multi-intent detection is more common in real-world scenarios, especially in real-time dialogue systems. Early research on multi-intent detection (Kim et al., 2017a; Gangadharaiah, 2019) attempted to apply traditional methods based on Convolutional Neural Networks or Recurrent Neural Networks , which were effective for capturing sequential patterns in input data. However, these models struggled with the complexity and ambiguity inherent in handling multiple intents simultaneously.

To address these challenges, Qin et al. (2020) proposed the Adaptive Graph-Interactive Framework, which incorporates graph-based interactions to model relationships between intents and slots more effectively. This approach was further extended to a non-autoregressive model by Qin et al. (2021a), improving the efficiency of multi-intent detection while reducing the limitations imposed by autoregressive structures. The non-autoregressive model offers a significant advantage in terms of computational efficiency, enabling real-time processing for multi-intent tasks.

In another approach, Cai et al. (2022) extends the JointBERT framework, which was initially designed for single-intent tasks, to handle multi-intent scenarios. It effectively addresses the shared-intent problem, where multiple intents may share common slot labels, by explicitly mapping the relationship between slots and their corresponding intents.

Recognizing the importance of accurately predicting the number of intents in a given utterance, Chen et al. (2022c) developed a novel threshold-free framework. This framework first predicts the number of intents present in the input before proceeding to predict the specific intents. This approach provides a more efficient and reliable way to handle multi-intent tasks, as it eliminates the need for predefined thresholds, which can vary across different datasets or domains.

2.2 Contrastive Learning

Contrastive Learning has gained widespread application in Natural Language Understanding tasks due to the challenges of data scarcity and the diversity of expressions in real-world language. Recent studies have demonstrated the effectiveness of CL in enhancing NLU performance (Gunel et al., 2020;

Hou et al., 2021; Yehudai et al., 2023) Specifically, for multi-intent NLU tasks, Vulić et al. (2022) proposed a strategy that adapts a general sentence encoder into a task-specific one for multi-intent data by leveraging contrastive learning. This approach improves the encoder's ability to distinguish between multiple intents in a single input.

In a similar vein, Tu et al. (2023) introduced a novel bidirectional joint model trained using supervised contrastive learning and self-distillation. This model effectively utilizes both intent and slot features, allowing them to complement each other, and improving the overall performance of multi-intent NLU tasks. The integration of supervised CL with self-distillation enables the model to learn richer representations, enhancing its ability to handle complex multi-intent scenarios.

3 Model Design

The LCAN model is designed around a label-aware attention mechanism, which enhances semantic discrimination through contrastive supervision. It further introduces a collaborative attention module to enable bidirectional interaction between intents and slots for efficient joint modeling. As illustrated in Figure 2, the overall architecture consists of four main components: a text encoder, a label-aware attention mechanism, a contrastive learning module, and intent-slot decoders.

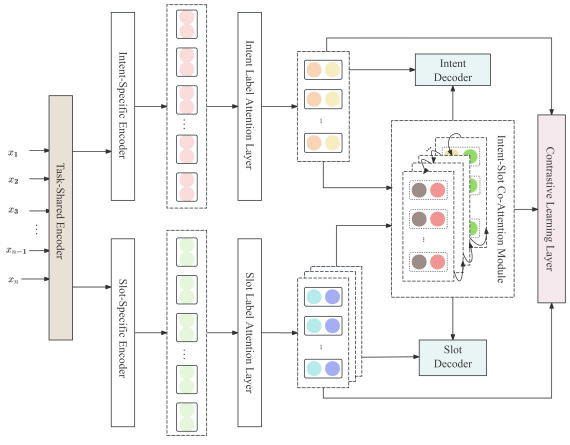

Figure 2: Illustration of the architecture of our joint model LCAN.

3.1 Text Encoder

Following previous work (Qin et al., 2020; Song et al., 2022),we employ a shared encoder and task-specific encoders.

3.1.1 Shared Encoder

We first use a shared encoder to extract general features. The shared encoder is based on RoBERTa to obtain contextual representations for each word. For an input sentence $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, where n is the number of words, the shared encoder produces the following context-aware word embeddings $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$, where each $\mathbf{h}_i$ is the contextualized embedding of word x_i:

$$\mathbf{H} = \text{RoBERTa}(\mathbf{x}) \tag{1}$$

3.1.2 Task-Specific Encoders

To refine the representations for intent recognition and slot filling, we introduce two independent task-specific BiLSTM encoders. The intent-specific encoder captures global semantic features, while the slot-specific encoder focuses on fine-grained word-level features.

The input sequence $\mathbf{H}$ is passed through the Intent BiLSTM to get the refined representation for intent recognition, denoted as $\mathbf{H}_{\text{intent}}$, and through the Slot BiLSTM to get the refined representation for slot filling, denoted as $\mathbf{H}_{\text{slot}}$:

$$\mathbf{H}_{\text{intent}} = \text{BiLSTM}_{\text{intent}}(\mathbf{H}) \tag{2}$$

$$\mathbf{H}_{\text{slot}} = \text{BiLSTM}_{\text{slot}}(\mathbf{H}) \tag{3}$$

3.2 Label-Aware Attention

To extract label-specific contextual representations from the input sequence, LCAN introduces a multi-level label attention mechanism for both intent and slot labels. This module learns individual attention distributions for each label, generating semantic vectors that serve as prior knowledge for joint modeling.

3.2.1 Intent Label Attention

The goal of intent label attention is to obtain the semantic representation of each token in the input sentence for every potential intent label. Initially, the model uses a specialized encoder to extract token-level representations $\mathbf{H}_{\text{intent}} \in \mathbb{R}^{d_e \times n}$, where n is the number of tokens, and d_e is the embedding dimension. A query matrix $\mathbf{Q}^{\text{I}} \in \mathbb{R}^{|\text{L}^{\text{I}}| \times d_a}$, where $|\text{L}^{\text{I}}|$ is the number of intent labels, is initialized with label embeddings. This query matrix is projected into the attention space, and attention weights are computed as:

$$\mathbf{A}^{\text{I}} = \text{softmax}\big(\mathbf{Q}^{\text{I}} \times \tanh\big(\mathbf{D}^{\text{I}} \times \mathbf{H}_{\text{intent}}\big) \in \mathbb{R}^{|\text{L}^{\text{I}}| \times n} \tag{4}$$

where, the softmax function normalizes across the row dimension, ensuring that each label's weights sum to one, and $\mathbf{D}^{\mathrm{I}} \in \mathbb{R}^{d_a \times d_e}$. The sentence-level representation for each intent label is calculated via a weighted sum of token embeddings:

$$\mathbf{V}^{\mathrm{I}} = \mathbf{H}_{\text{Intent}} \times \left(\mathbf{A}^{\mathrm{I}}\right)^{T} \in \mathbb{R}^{d_e \times |\mathrm{L}^{\mathrm{I}}|} \tag{5}$$

3.2.2 Slot Label Attention

In contrast to intent labels, slot labels usually have more complex hierarchical structures, we introduce a multi-layer slot label attention mechanism. At each layer $\mathbf{H}_{\text{slot}} \in \mathbb{R}^{d_e \times n}$, the model extracts token-level representations $\mathbf{Q}^{\mathrm{S},k} \in \mathbb{R}^{|\mathrm{L}^{\mathrm{S},k}| \times d_a}$ from a slot-specific encoder. A query matrix $\mathbf{D}^{\mathrm{S},k} \in \mathbb{R}^{d_a \times d_e}$ is used to calculate attention weights, which are projected into the attention space as follows:

$$\mathbf{A}^{\mathrm{S},k} = \mathrm{softmax}\big(\mathbf{Q}^{\mathrm{S},k} \times \tanh\big(\mathbf{D}^{\mathrm{S},k} \times \mathbf{H}_{\text{slot}}\big) \tag{6}$$

After calculating the attention for each slot label at level k, the model computes the sentence representation for all slot labels at that level as:

$$\mathbf{V}^{\mathrm{S},k} = \mathbf{H}_{\text{Slot}} \times \left(\mathbf{A}^{\mathrm{S},k}\right)^{T} \in \mathbb{R}^{d_e \times |\mathrm{L}^{\mathrm{S},k}|} \tag{7}$$

3.2.3 Fine-Grained Semantic Enhancement

To fully leverage the semantic information from slot labels at higher layers, the model enhances the fine-grained slot label representations by incorporating semantic information from coarser labels. Specifically, the model applies a feed-forward network followed by a sigmoid activation function to estimate the prediction probability for each coarse-level slot label:

$$p_j^{\mathrm{S},k-1} = \sigma(\mathbf{w}_j^{\mathrm{S},k-1} \cdot \mathbf{v}_j^{\mathrm{S},k-1}) \tag{8}$$

where $w_j^{\mathrm{S},k-1}$ represents the weight for each coarse-level slot label class. These probabilities are concatenated and projected into a unified semantic vector:

$$\mathbf{Z}^{\mathrm{S},k-1} = \mathbf{Z}^{\mathrm{S},k-1} \cdot \mathbf{p}^{\mathrm{S},k-1} \in \mathbb{R}^{d_p} \tag{9}$$

This vector is then concatenated with each fine-grained slot label vector to form an enhanced fine-grained label vector:

$$\mathbf{v}_j^{\mathrm{S},k} \longleftarrow \mathbf{v}_j^{\mathrm{S},k} \oplus \mathbf{z}^{\mathrm{S},k-1} \tag{10}$$

Finally, the updated slot label representation at the $k-th$ layer is:

$$\mathbf{V}^{\mathrm{S},k} = \left[\mathbf{v}_1^{\mathrm{S},k}, \mathbf{v}_2^{\mathrm{S},k}, \ldots, \mathbf{v}_{|\mathrm{L}^{\mathrm{S},k}|}^{\mathrm{S},k}\right] \in \mathbb{R}^{(d_e+d_p) \times |\mathrm{L}^{\mathrm{S},k}|} \tag{11}$$

Through this hierarchical enhancement mechanism, the model refines the fine-grained slot label representations by integrating information from coarser slot labels, which improves the overall performance of slot filling tasks.

3.2.4 Intent-Slot Co-attention

We propose an intent-slot co-attention mechanism that enables fine-grained bidirectional interaction via label-specific representations.

The module takes as input a set of label-specific representations generated by the label attention layer, including the semantic representation matrix of intent labels $\mathbf{Q}_1 = \mathbf{V}^{\mathrm{I}} \in \mathbb{R}^{d_e \times |\mathrm{L}^{\mathrm{I}}|}$, each layer of slot labels is represented by semantic embeddings $\mathbf{Q}_2 = \mathbf{V}^{\mathrm{S},1}, \mathbf{Q}_3 = \mathbf{V}^{\mathrm{S},2}, \ldots, \mathbf{Q}_{l+1} = \mathbf{V}^{\mathrm{S},l}$, along with a soft slot embedding matrix for each token $\mathbf{Q}_{l+2} = \mathbf{S} \in \mathbb{R}^{d_s \times n}$. These inputs collectively form a semantic hierarchical sequence for modeling label-level interactions via the co-attention mechanism.

Since the input embeddings may differ in dimensionality, all representations are first projected into a unified semantic space to enable cross-layer interaction. To achieve this, we introduce separate forward and backward projection matrices for each input $\mathbf{Q}_t \in \mathbb{R}^{d_t \times m_t}$, mapping them into a shared representation space of dimension d:

$$\overrightarrow{\mathbf{Q}}_t = \overrightarrow{\mathbf{W}}_t \mathbf{Q}_t, \overleftarrow{\mathbf{Q}}_t = \overleftarrow{\mathbf{W}}_t \mathbf{Q}_t \tag{12}$$

where $\overrightarrow{\mathbf{Q}}_t$ and $\overleftarrow{\mathbf{Q}}_t$ serve as the basis for semantic propagation: the former supports forward propagation (slot-to-intent), while the latter is used for backward propagation (intent-to-slot). At each step, a bilinear attention module computes a correlation matrix between adjacent layers:

$$\mathbf{C}_t = \mathbf{Q}_{t-1}^{t} \mathbf{X}_t \mathbf{Q}_t \in \mathbb{R}^{m_{t-1} \times m_t} \tag{13}$$

where $\mathbf{X}_t \in \mathbb{R}^{d_{t-1} \times d_t}$ is a learnable attention weight matrix capturing inter-label similarity. Semantic propagation is then performed in both directions. For forward propagation (slot-to-intent), the update is:

$$\overrightarrow{\mathbf{H}}_t = \begin{cases} \tanh\left(\overrightarrow{\mathbf{Q}}_{t-1} \mathbf{C}_t + \overrightarrow{\mathbf{Q}}_t\right), & t = 2 \\ \tanh\left(\overrightarrow{\mathbf{H}}_{t-1} \mathbf{C}_t + \overrightarrow{\mathbf{Q}}_t\right), & t > 2 \end{cases} \tag{14}$$

Conversely, backward propagation (intent-to-slot) is defined as:

$$\overleftarrow{\mathbf{H}}_t = \begin{cases} \tanh\left(\overleftarrow{\mathbf{Q}}_{t+1} \mathbf{C}_{t+1}^{T} + \overleftarrow{\mathbf{Q}}_t\right), & t = l+1 \\ \tanh\left(\overleftarrow{\mathbf{H}}_{t+1} \mathbf{C}_{t+1}^{T} + \overleftarrow{\mathbf{Q}}_t\right), & t < l+1 \end{cases} \tag{15}$$

This enables intent-level semantic cues to flow back to the slot layer, supporting more accurate slot predictions through joint label awareness. The dual-path propagation forms a closed loop of semantic communication, improving contextual understanding across labels.

After propagation, the model extracts final label representations for decoding. Specifically, the first-layer output from backward propagation is used as the intent representation, while the final output of the forward path is used for slot decoding:

$$\mathbf{R}_{\mathrm{I}}^{final} = \overline{\mathbf{H}}_1 \in \mathbb{R}^{d \times |\mathrm{L}^{\mathrm{I}}|} \tag{16}$$

$$\mathbf{R}_{\mathrm{S}}^{final} = \overline{\mathbf{H}}_{l+2} \in \mathbb{R}^{d \times n} \tag{17}$$

3.3 Contrastive Learning

While label-aware attention captures sentence-level semantics, semantic overlap persists among similar labels in multi-intent/slot scenarios. LCAN addresses this through a label-aware contrastive learning module (Figure 3), structuring the embedding space via contrastive supervision. Following supervised contrastive principles, it creates positive pairs and negative pairs, enforcing intra-class compactness and inter-class separation. The framework extends to three semantic levels: intent labels, slot labels, and cross-task intent-slot pairs, establishing multi-granular contrastive supervision that enhances discriminative power while mitigating feature confusion.

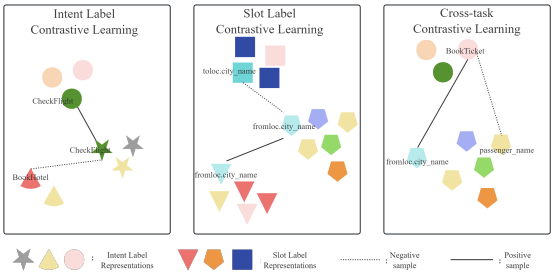

Figure 3: Diagram of the Contrastive Learning Module.

3.3.1 Intent Label Contrastive Learning

Intent detection is a multi-label classification task at the sentence level. Semantic similarity between intent labels is common—especially when multiple intents co-occur within a user query. LCAN first extracts per-label semantic vectors via label-aware attention:

$$\mathbf{V}^{\mathrm{I}} = [\,\mathbf{v}_1^{\mathrm{I}}, \mathbf{v}_2^{\mathrm{I}}, \ldots, \mathbf{v}_{|\mathrm{L}^{\mathrm{I}}|}^{\mathrm{I}}] \in \mathbb{R}^{d_e \times |\mathrm{L}^{\mathrm{I}}|} \tag{18}$$

Positive pairs are sampled from embeddings of the same intent label across different sentences, while negative pairs are drawn from different labels. The model is trained using the InfoNCE loss:

$$\mathcal{L}_{\mathrm{i-cl}} = -\,log \frac{\exp\left(sim(v, v^+)/\tau\right)}{\Sigma_{v^- \in \mathcal{N}} \exp\left(sim(v, v^-)/\tau\right)} \tag{19}$$

where $sim()$ denotes cosine similarity, τ is a temperature parameter, and $\mathcal{N}$ is the set of negatives. This loss pulls together embeddings of the same intent and pushes apart others, enhancing robustness in multi-label scenarios.

3.3.2 Slot Label Contrastive Learning

Slot filling involves a larger label space and finer semantic granularity. Overlapping meanings among hierarchical slot labels further complicate representation learning. LCAN generates multi-layer slot label embeddings $\mathbf{V}^{\mathrm{S},k} \in \mathbb{R}^{d_k \times |\mathrm{L}^{\mathrm{S},k}|}$ via hierarchical attention computed following Equation 11.

For each layer k, contrastive learning is applied similarly:

$$\ell_i = \log \frac{exp\left(sim(v_i, v_i^+)/\tau\right)}{\sum_{v_j^- \in \mathcal{N}} exp\left(sim(v_i, v_j^-)/\tau\right)} \tag{20}$$

$$\mathcal{L}_{\mathrm{s},k} = -\frac{1}{|\mathrm{L}^{\mathrm{S},k}|} \sum_{i=1}^{|\mathrm{L}^{\mathrm{S},k}|} \ell_i \tag{21}$$

This contrastive loss encourages the model to differentiate between semantically similar slot labels.

$$\mathcal{L}_{\mathrm{s-cl}} = \Sigma_{k=1}^{K} \mathcal{L}_{\mathrm{s},k} \tag{22}$$

3.3.3 Cross-Task Contrastive Learning

Intent and slot labels are closely correlated. To align semantic spaces across tasks, LCAN incorporates cross-task contrastive learning, enforcing similarity between semantically aligned intent-slot label pairs. Given an intent embedding $\mathbf{v}_i^{\mathrm{I}}$ and a related slot embedding $\mathbf{v}_j^{\mathrm{S},k}$, a contrastive loss is defined as:

$$\mathcal{L}_{cross} = -\sum_{i,j \in \mathcal{P}} \log \frac{exp\left(sim(\mathbf{v}_i^{\mathrm{I}}, \mathbf{v}_j^{\mathrm{S},k})/\tau\right)}{\sum_{i,j^- \in \mathcal{N}} exp\left(sim(\mathbf{v}_i^{\mathrm{I}}, \mathbf{v}_{j^-}^{\mathrm{S},k})/\tau\right)} \tag{23}$$

3.4 Intent and Slot Decoders

The final stage of the model involves decoding the intent and slot labels. Both the intent and slot decoders are based on a multi-label classification

approach for intent recognition and a sequence labeling approach for slot filling.

3.4.1 Intent Decoder

Intent detection is modeled as a multi-label classification task, where a single utterance may correspond to multiple semantic intents. To handle this, LCAN employs a sigmoid-based multi-label classifier.To train the model, a binary cross-entropy loss is computed independently for each label and summed across all labels:

$$\mathcal{L}_i = -\sum_{j=1}^{\lfloor \mathrm{L^I} \rfloor} \left[y_j^{\mathrm{I}} \, log\,(\hat{y}_j^{\mathrm{I}}) + (1 - y_j^{\mathrm{I}}) \, log\,(1 - \hat{y}_j^{\mathrm{I}}) \right] \quad (24)$$

where y_j^{I} denotes the ground-truth label for the $i-th$ intent. This objective supports robust learning for multi-intent utterances.

3.4.2 Slot Decoder

Slot filling is framed as a sequence labeling task, where each token in the utterance is assigned a corresponding slot label. To account for contextual dependencies among slot tags, LCAN adopts a hybrid decoding approach using a linear classifier followed by a Conditional Random Field .

To model label transitions and enforce sequential consistency, a CRF layer is applied over the emission scores. Given a sequence of slot labels $y^{\mathrm{S}} = y_1^{\mathrm{S}}, y_2^{\mathrm{S}}, ..., y_n^{\mathrm{S}}$, the sequence-level score is defined as:

$$score(x, y^{\mathrm{S}}) = \Sigma_{i=1}^{n} \left(emission_{i, y_i^{\mathrm{S}}} + T_{y_{i-1}^{\mathrm{S}}, y_i^{\mathrm{S}}} \right) \quad (25)$$

where $emission_{i, y_i^{\mathrm{S}}}$ denotes the transition score from the previous label to the current label.

The CRF is trained by maximizing the log-likelihood of the correct label sequence, which is equivalent to minimizing the negative log-likelihood:

$$\mathcal{L}_s = -score(\mathbf{x}, y^{\mathrm{S}}) + log\,\Sigma_{\hat{\mathbf{y}} \in \mathcal{Y}}\, exp\,(\,score(\mathbf{x}, \hat{\mathbf{y}})) \quad (26)$$

The denominator sums over all possible label sequences and is computed efficiently via the forward algorithm. During inference, the Viterbi algorithm is used to find the most probable label sequence.

3.5 Overall Training Objective

The LCAN model is trained using a joint objective function that combines two major components: 1)Task-specific supervised losses, including a multi-label classification loss for intent detection and a sequence labeling loss for slot filling; 2)Label-aware contrastive losses, designed to enhance semantic separability among intent and slot representations.

The overall loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_i + \lambda_0 \mathcal{L}_s + \lambda_1 \mathcal{L}_{i-cl} + \lambda_2 \mathcal{L}_{s-cl} + \lambda_3 \mathcal{L}_{cross} \quad (27)$$

This flexible multi-objective formulation enables the model to balance structural supervision and semantic discriminability, improving performance and generalization across both tasks.

4 Experiments Setup

4.1 Datasets

The experiments are conducted on two standard benchmark datasets: MixATIS (Qin et al., 2020) and MixSNIPS (Qin et al., 2020). MixATIS is a dataset primarily focusing on the airline travel domain, containing multiple intents like flight booking, flight inquiry, and time-based queries. The MixSNIPS dataset, on the other hand, covers multiple domains such as weather, music, restaurants, and movies, with a more conversational and informal style of language.

4.2 Implementation Details

We use RoBERTa-base as the shared encoder to extract general semantic features from input sequences. Two task-specific encoders based on BiLSTM are employed: one for intent recognition and one for slot filling. The label-aware attention mechanism is introduced to refine the interaction between intent and slot labels. Additionally, the model incorporates a contrastive learning strategy to enhance semantic differentiation between similar or overlapping labels. The AdamW optimizer (Loshchilov and Hutter, 2017) is used for model training with an initial learning rate of 0.001. The batch size is set to 32, and the maximum number of training epochs is 50. Early stopping is applied to prevent overfitting. We perform hyperparameter tuning using random search on the validation set.

4.3 Baselines

To comprehensively evaluate the effectiveness of the proposed LCAN model on multi-intent detection and slot filling tasks, we conducted systematic comparative experiments on MixATIS and MixSNIPS. The evaluation includes several representative baseline models, ranging from traditional joint modeling approaches to recent frameworks enhanced by graph structures and atten-

tion mechanisms:(1)GL-GIN (Qin et al., 2021b): A non-autoregressive model with global-local graph interaction for efficient intent-slot coordination.(2)SDJN (Chen et al., 2022b): Integrates self-distillation and multi-instance learning to enhance word-level intent-slot alignment.(3)Co-guiding Net (Xing and Tsang, 2022): Employs heterogeneous graph attention and contrastive learning for bidirectional task interaction.(4)TFMN (Chen et al., 2022c): Multi-task framework with auxiliary intent count prediction, eliminating threshold dependencies.(5)BiSLU (Tu et al., 2023): Combines supervised contrastive learning with self-distillation for explicit intent-slot collaboration.(6)MISCA (Pham et al., 2023): Joint attention architecture enabling implicit intent-slot information exchange without graph construction.(7)Uni-MIS (Yin et al., 2024): Models the multi-intent SLU as a multi-view intent-slot interaction.

5 Results and Analysis

5.1 Main Results

Experimental results are presented in Table 1. On both the MixATIS and MixSNIPS datasets, the proposed LCAN model demonstrates a clear overall performance advantage in joint multi-intent modeling tasks.

On the MixATIS dataset, LCAN achieves a joint accuracy of 59.5%, outperforming MISCA and BiSLU by 1.4% and 8.0%, respectively. This confirms its effectiveness in handling complex hierarchical slot labels and diverse intents in the airline domain. Notably, LCAN achieves an F1-score of 92.4% on slot filling, 3.0% higher than BiSLU, indicating that its label-aware attention mechanism effectively captures fine-grained slot dependencies. Although its intent accuracy is slightly lower than BiSLU, the incorporation of cross-task contrastive learning strengthens semantic alignment between intents and slots, leading to an improvement in overall joint accuracy.

On the open-domain MixSNIPS dataset, LCAN achieves strong intent detection performance, benefiting from contrastive learning to alleviate semantic overlap among multiple intents and to enhance global consistency across tasks. While its slot F1-score is slightly lower than BiSLU, LCAN still achieves superior joint accuracy, highlighting its advantage in optimizing intent-slot interactions. Compared to earlier models, LCAN improves joint accuracy by 7.8% over Co-guiding Net, further validating its robustness in multi-domain, multi-intent scenarios.

Table 1: Obtained results

Model	MixATIS			MixSNIPS		
	Intent (Acc.)	Slot (F1)	Overall (Acc.)	Intent (Acc.)	Slot (F1)	Overall (Acc.)
GL-GIN	76.3	88.3	43.5	95.6	94.9	75.4
SDJN	77.1	88.2	44.6	96.5	94.4	75.7
Co-guiding	79.1	89.8	51.3	97.7	95.1	77.5
TFMN	79.8	88.0	50.2	97.7	96.4	84.7
BiSLU	**81.5**	89.4	51.5	97.8	**97.2**	**85.4**
MISCA	80.6	90.0	58.1	97.6	96.1	83.1
Uni-MIS	78.5	88.3	52.5	97.2	96.4	83.4
LCAN (Ours)	81.3	**92.4**	**59.5**	**98.7**	96.9	85.3

5.2 Ablation Results

To evaluate the contribution of different contrastive learning modules, we conduct an ablation study by removing various contrastive learning modules. The results are shown in Table 2, and the ablated variants include: w/o Intent CL, w/o Slot CL, w/o Cross CL, and w/o CL.

Table 2: Ablation Study on Contrastive Learning Module

Variant	MixATIS			MixSNIPS		
	Intent (Acc.)	Slot (F1)	Overall (Acc.)	Intent (Acc.)	Slot (F1)	Overall (Acc.)
w/o CL	80.6	90.0	58.3	97.6	96.1	83.1
w/o Intent CL	79.9	91.8	58.9	97.7	96.5	84.8
w/o Slot CL	80.5	89.6	58.7	98.5	94.9	84.3
w/o Cross CL	**81.6**	91.7	58.8	**98.8**	95.9	84.8
LCAN	81.3	**92.4**	**59.5**	98.7	**96.9**	**85.3**

Removing all contrastive learning modules leads to a decrease of 1.2% and 2.2% in joint accuracy on MixATIS and MixSNIPS, respectively. This confirms the effectiveness of contrastive learning in enhancing semantic consistency between intents and slots, thereby improving overall performance.On MixATIS, intent accuracy drops 1.4%, indicating its primary focus on intent representation. For MixSNIPS , it proves vital in open-domain intent-dominant scenarios.Causes significant slot F1 declines, highlighting its importance for fine-grained slot recognition. The 0.8% intent accuracy drop on MixATIS suggests improved slot distinction indirectly benefits intent inference.Removal increases

MixATIS intent accuracy due to reduced overfitting but lowers slot F1 and joint accuracy, demonstrating its balance between task interaction and semantic alignment. On MixSNIPS, joint accuracy drops 0.5% despite minor intent gains. In summary, the intent, slot, and cross contrastive learning modules contribute to the LCAN model from global intent optimization, local slot representation, and semantic coordination, respectively, leading to improved multi-intent joint modeling.

To evaluate the contributions of label-aware attention structure, we design three simplified model variants by removing specific components of the label-aware attention structure. The ablated variants include: w/o Co-attention, w/o Slot Label Attention, and w/o Intent Label Attention.

Table 3: Ablation Study on Label Attention Mechanism

Variant	MixATIS			MixSNIPS		
	Intent (Acc.)	Slot (F1)	Overall (Acc.)	Intent (Acc.)	Slot (F1)	Overall (Acc.)
w/o Co-Attn	79.9	89.2	50.5	96.9	95.9	80.1
w/o SL Attn	79.6	91.2	56.1	**98.7**	96.1	81.4
w/o IL Attn	78.7	91.6	55.3	97.2	96.2	82.0
LCAN	**81.3**	**92.4**	**59.5**	**98.7**	**96.9**	**85.3**

As shown in Table 3, removing the co-attention mechanism leads to decrease of 9.0% and 5.2% in joint accuracy on MixATIS and MixSNIPS, respectively. This confirms the effectiveness of co-attention mechanism in modeling semantic coordination and promoting collaborative decoding between the intents and entities. The absence of SL Attn prevents the model from establishing fine-grained semantic alignments between slot labels and the input context. This leads to a clear accuracy drop, especially on MixSNIPS (nearly 4%). Furthermore, the removal of the IL Attn leads to a substantial drop in intent classification accuracy, this suggests that relying solely on task-specific encoder outputs is insufficient to capture the semantic diversity of multiple intents.

5.3 Case Study

To further illustrate the effectiveness of LCAN, we conduct a case study using a representative multi-intent example from the MixATIS dataset: "Check flights from Beijing to Shanghai and book a ticket" . Traditional models exhibit blurred intent boundaries at conjunctions, while RoBERTa-base fails to disambiguate slots near "and". As shown in Table 4, BiSLU improves intent separation but retains coarse slot granularity.This demonstrates LCAN's superiority in disentangling interconnected intents and maintaining slot-label consistency.

Table 4: Recognition Results for a Multi-Intent Example

Model	Recognized Intents	Slot Filling Results
RoBERTa-base	CheckFlight, BookTicket	from: Beijing; to: Shanghai and book; ticket: [Unrecognized]
BiSLU	CheckFlight, BookTicket	from: Beijing; to: Shanghai; ticket: ticket
LCAN	CheckFlight, BookTicket	from: Beijing; to: Shanghai; ticket: book a ticket

Table 4 shows the recognition results of three models under this input. While all models successfully detect the two intents, there are clear differences in slot filling accuracy and boundary clarity. RoBERTa-base fails to distinguish between the intents around the conjunction, leading to incomplete slot extraction. BiSLU improves intent separation through bidirectional learning and contrastive strategies but still shows coarse-grained slot interpretation. LCAN, with its label-aware mechanism, accurately aligns slot information with each intent, demonstrating superior semantic parsing in multi-intent scenarios.

6 Conclusion

We propose LCAN, a novel joint model addressing multi-intent recognition and slot filling challenges in task-oriented dialogue systems. LCAN integrates label-aware attention for precise input-label alignment and contrastive learning to mitigate semantic ambiguity among overlapping labels. Evaluations on MixATIS and MixSNIPS datasets demonstrate LCAN's superiority over state-of-the-art models, particularly in complex multi-intent scenarios requiring nuanced intent-slot interactions. Ablation studies confirm the critical roles of contrastive learning and label-aware attention. While LCAN achieves robust performance, future work could explore hierarchical attention mechanisms and optimized contrastive strategies for improved generalization to unseen, highly diverse dialogues. The framework establishes a scalable foundation for advancing joint modeling in multi-intent dialogue systems.

Limitations

It should also be emphasized that LCAN integration of label-aware attention mechanisms and contrastive learning modules may substantially increase both parameter count and computational complexity. More critically, the efficacy of contrastive learning is contingent upon the design of positive/negative sample pairs—where semantic overlaps between intents and slots in multi-intent scenarios can induce sample selection bias. Furthermore, the label-level focus of contrastive learning might neglect finer-grained semantic distinctions at the word or phrase level. Despite LCAN consistent outperformance of existing models, we posit that incorporating fine-grained semantic similarity computation could yield additional performance gains.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (62074055, 62174038, 62374047), and State Key Lab of Processors, Institute of Computing Technology, CAS (CLQ 202407).

References

Fengyu Cai, Wanhao Zhou, Fei Mi, and Boi Faltings. 2022. Slim: Explicit slot-intent mapping with bert for joint multi-intent detection and slot filling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7607–7611. IEEE.

Dongsheng Chen, Zhiqi Huang, Xian Wu, Shen Ge, and Yuexian Zou. 2022a. Towards joint intent detection and slot filling via higher-order attention. In *IJCAI*, pages 4072–4078.

Guanhua Chen, Yutong Yao, Derek F Wong, and Lidia S Chao. 2024. A two-stage prediction-aware contrastive learning framework for multi-intent nlu. *arXiv preprint arXiv:2405.02925*.

Lisong Chen, Peilin Zhou, and Yuexian Zou. 2022b. Joint multiple intent detection and slot filling via self-distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7612–7616. IEEE.

Lisung Chen, Nuo Chen, Yuexian Zou, Yong Wang, and Xinzhong Sun. 2022c. A transformer-based threshold-free framework for multi-intent nlu. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7187–7192.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Lizhi Cheng, Wenmian Yang, and Weijia Jia. 2023a. A scope sensitive and result attentive model for multi-intent spoken language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12691–12699.

Xuxin Cheng, Wanshi Xu, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023b. Towards spoken language understanding via multi-level multi-grained contrastive learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 326–336.

Rashmi Gangadharaiah. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Yutai Hou, Yongkui Lai, Cheng Chen, Wanxiang Che, and Ting Liu. 2021. Learning to bridge metric spaces: Few-shot joint learning of intent detection and slot filling. *arXiv preprint arXiv:2106.07343*.

Byeongchang Kim, Seonghan Ryu, and Gary Geunbae Lee. 2017a. Two-stage multi-intent detection for spoken language understanding. *Multimedia Tools and Applications*, 76:11377–11390.

Young-Bum Kim, Sungjin Lee, and Karl Stratos. 2017b. Onenet: Joint domain, intent, slot prediction for spoken language understanding. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 547–553. IEEE.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

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

Zhiyuan Ma, Jiwei Qin, Meiqi Pan, Song Tang, Jinpeng Mi, and Dan Liu. 2024. Promoting unified generative framework with descriptive prompts for joint multi-intent detection and slot filling. *Electronics*, 13(6):1087.

Thinh Pham, Chi Tran, and Dat Quoc Nguyen. 2023. Misca: A joint model for multiple intent detection and slot filling with intent-slot co-attention. *arXiv preprint arXiv:2312.05741*.

Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021a. A co-interactive transformer for joint slot filling and intent detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8193–8197. IEEE.

Libo Qin, Fuxuan Wei, Tianbao Xie, Xiao Xu, Wanxiang Che, and Ting Liu. 2021b. Gl-gin: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling. *arXiv preprint arXiv:2106.01925*.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. Agif: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. *arXiv preprint arXiv:2004.10087*.

Mengxiao Song, Bowen Yu, Li Quangang, Wang Yubin, Tingwen Liu, and Hongbo Xu. 2022. Enhancing joint multiple intent detection and slot filling with global intent-slot co-occurrence. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 7967–7977.

Nguyen Anh Tu, Hoang Thi Thu Uyen, Tu Minh Phuong, and Ngo Xuan Bach. 2023. Joint multiple intent detection and slot filling with supervised contrastive learning and self-distillation. In *ECAI 2023*, pages 2370–2377. IOS Press.

Ivan Vulić, Iñigo Casanueva, Georgios Spithourakis, Avishek Mondal, Tsung-Hsien Wen, and Paweł Budzianowski. 2022. Multi-label intent detection via contrastive task specialization of sentence encoders. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7544–7559.

Jixuan Wang, Kai Wei, Martin Radfar, Weiwei Zhang, and Clement Chung. 2021. Encoding syntactic knowledge in transformer encoder for intent detection and slot filling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13943–13951.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling. *arXiv preprint arXiv:2010.02693*.

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*.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 ieee workshop on automatic speech recognition and understanding*, pages 78–83. IEEE.

Asaf Yehudai, Matan Vetzler, Yosi Mass, Koren Lazar, Doron Cohen, and Boaz Carmeli. 2023. Qaid: Question answering inspired few-shot intent detection. *arXiv preprint arXiv:2303.01593*.

Shangjian Yin, Peijie Huang, and Yuhong Xu. 2024. Uni-mis: United multiple intent spoken language understanding via multi-view intent-slot interaction. In *AAAI*, pages 19395–19403.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*, volume 16, pages 2993–2999.

Xianwei Zhuang, Xuxin Cheng, and Yuexian Zou. 2024. Towards explainable joint models via information theory for multiple intent detection and slot filling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19786–19794.