Pairwise Prompt-Based Tuning with Parameter Efficient Fast Adaptation for Generalized Zero-Shot Intent Detection

Xiaotong Zhang, Qianru Zhou, Han Liu*, Hong Yu
School of Software Technology, Dalian University of Technology
zxt.dut@hotmail.com, zhouqianru.dut@gmail.com,
liu.han.dut@gmail.com, hongyu@dlut.edu.cn

Abstract

Generalized zero-shot intent detection (GZID) aims to recognize the labels of utterances from both seen and unseen intents by utilizing the knowledge learned from seen intents. Enhancing the generalization ability from seen intents to unseen intents is a key challenge in the GZID setting. Existing methods attempt to tackle this challenge by distinguishing unseen intents from seen intents or focusing on enhancing the model discriminability. However, the challenge is not solved substantially as they ignore to promote the representation learning ability of the model itself and neglect to strengthen the model adaptability to new tasks, resulting in overfitting on the seen intents. In this paper, we propose a pairwise prompt-based tuning model with parameter efficient fast adaptation which involves two training steps. In the first step, we leverage hybrid contrastive learning in discriminant space and masked language modeling to make predictions at both sentence and token levels, which can enhance the model discriminability and representation learning ability respectively. In the second step, we design a pipeline for generating and filtering unseen data by only providing unseen intent labels, and utilize parameter-efficient fine-tuning to quickly adapt to unseen intents. Experiments on four intent detection datasets demonstrate that our two-step training method has better comprehension and generalization capabilities.

1 Introduction

Conversational systems such as social robots and voice assistants emerge almost in every digital device (Seminck, 2023), which need to comprehend human intent to provide appropriate responses or execute corresponding instructions. However, with the expansion of conversational systems, the dialogue topics are no longer limited to predefined domains. To recognize the fast-emerging intents, collecting the annotated data and retraining the intent detection model are resource-intensive. To solve this issue, zero-shot intent detection is proposed to detect unseen intents with the help of learned knowledge from seen intents. In standard zero-shot learning setting (Chang et al., 2008; Xian et al., 2019), the test set only contains data in unseen classes, making it impractical in real applications. Approaches designed under this circumstance is hard to generalize to unseen classes, i.e., classifying most test data into seen classes, which motivates generalized zero-shot intent detection (GZID) that classify utterances in both seen and unseen classes during inference (Socher et al., 2013; Atzmon and Chechik, 2019; Pourpanah et al., 2023).

The challenge of GZID lies in how to improve the generalization ability of identifying unseen intents. Recently, various GZID methods have been proposed. These include pre-partitioning approaches that first use a classifier such as LOF (Breunig et al., 2000) or PU (Su et al., 2021) to distinguish unseen intents from seen ones, and then classify utterances with zero-shot intent detection models. Some of these models such as Zero-Shot DNN (ZSDNN) (Kumar et al., 2017) learn the prototypes using neural network trained via a ranking loss. Others reconstruct the transformation matrices, like CapsNet (Xia et al., 2018) and ReCapsNet (Liu et al., 2019), or incorporate external knowledge, such as RIDE (Siddique et al., 2021). Additionally, the models above can further enhance their performance by leveraging a plugin called Class-Transductive Intent Representations (CTIR) (Si et al., 2021), a framework that incorporates unseen label names during training.

Another branch of approaches implement GZID via an end-to-end manner. The network called Learn to Adapt (LTA) (Zhang et al., 2022) simulates the training scenario of GZID by constructing virtual unseen categories from seen classes. This involves continually adjusting category prototypes and sample representations to obtain improved em-

[*] Corresponding author.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 917–929
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

beddings. SP RoBerta+template (Lamanov et al., 2022) utilizes templates to fuse utterances and intent labels, transforming intent recognition into a text entailment task (Yin et al., 2019), then it applies contrastive learning on multiple pairs of seen data and intent labels to enhance discriminability. AGCR (Liu et al., 2024) employs a generation-based method that utilizes a large pre-trained language model to produce pseudo-novel samples. The most representative samples are selected as category anchors, and labels are predicted based on similarity to these anchors. Despite its advantages, training from scratch on the generated unseen data can be resource-intensive and time-consuming.

Nevertheless, current models face two main issues. Firstly, the generalization ability of pre-partitioning methods highly relies on the performance in distinguishing unseen intents from seen intents in the first stage. Therefore, these methods sometimes struggle to adapt effectively to the GZID setting. Although some methods can directly detect unseen intents by skipping the first stage, the performance tends to be poor in the GZID setting. Secondly, while end-to-end methods are devoted to enhancing the model discriminability by minimizing classification errors in loss function, they often ignore to promote the representation learning as well as adaptability to new tasks, limiting the generalization ability when it comes to identifying unseen intents.

To address these issues, we propose a pairwise prompt-based tuning model with parameter efficient fast adaptation for GZID. This method involves two steps. In the first step, we concatenate the utterances and intent labels with a template, and conduct prompt-based tuning with two loss functions. The hybrid contrastive learning loss in discriminant space employs a more challenging negative sampling strategy by sampling both hard utterances and hard intents. The masked language modeling loss further leverages information from positive samples, enhancing the model's representation learning at the token level by predicting the random mask tokens. In the second step, we apply a parameter-efficient fine-tuning strategy called P-tuning (Liu et al., 2021) to quickly adapt the base model trained in the first step to unseen tasks. To accomplish this, we use a generative model to produce a small batch of unseen data, which is then filtered by a similarity model. Subsequently, P-tuning is applied to fit the unseen intents by freezing the base model's parameters from the first step and adding a small prompt encoder to facilitate rapid adaptation to unseen intents.

Our contributions can be outlined as follows. (1) We introduce a two-step training approach to address the challenge of generalizing to unseen intents in the GZID setting. (2) We propose a novel hybrid contrastive learning loss function in a discriminant space, coupled with masked language modeling loss targeting each positive pair of utterance and intent to make predictions at the sentence and token levels, respectively. (3) We devise a systematic process for acquiring unseen data through data generation and filtering, and employ a parameter-efficient fine-tuning strategy P-tuning to rapidly adapt the model to new intents with minimal parameter adjustments. (4) Experiments on four dialogue datasets demonstrate that our approach outperforms state-of-the-art baselines in the GZID setting.

2 Related Work

2.1 Generalized Zero-Shot Intent Detection

In the intent detection task, the generalized zero-shot learning (GZSL) methods can be broadly classified into three categories: transformation-based methods, compatibility-based methods and textual entailment based methods. Transformation-based methods such as CapsNet (Xia et al., 2018), ReCapsNet (Liu et al., 2019), and CTIR (Si et al., 2021) calculate inter-intent similarity based on word embeddings of intent labels. These methods leverage this similarity to effectively transform predictions from seen intents to unseen ones, establishing semantic connections between seen classes and unseen classes. Compatibility-based methods like ZSDNN (Kumar et al., 2017), LTA (Zhang et al., 2022), AGCR (Liu et al., 2024) strive to establish a unified semantic space for label names or category anchors and utterances through training on labeled data. This shared space facilitates the computation of similarity between unseen label names and test utterances, enabling effective prediction for unseen classes. Textual entailment-based methods (Lamanov et al., 2022) expand labels into coherent sentences, then concatenate them with corresponding samples to form sentence pairs. By inputting these pairs into PLMs, the similarity of sentence pairs is calculated using a linear classification head. This process transforms intent recognition into a text entailment task (Sun et al., 2022). Besides the GZID setting that has seen intents, some meth-

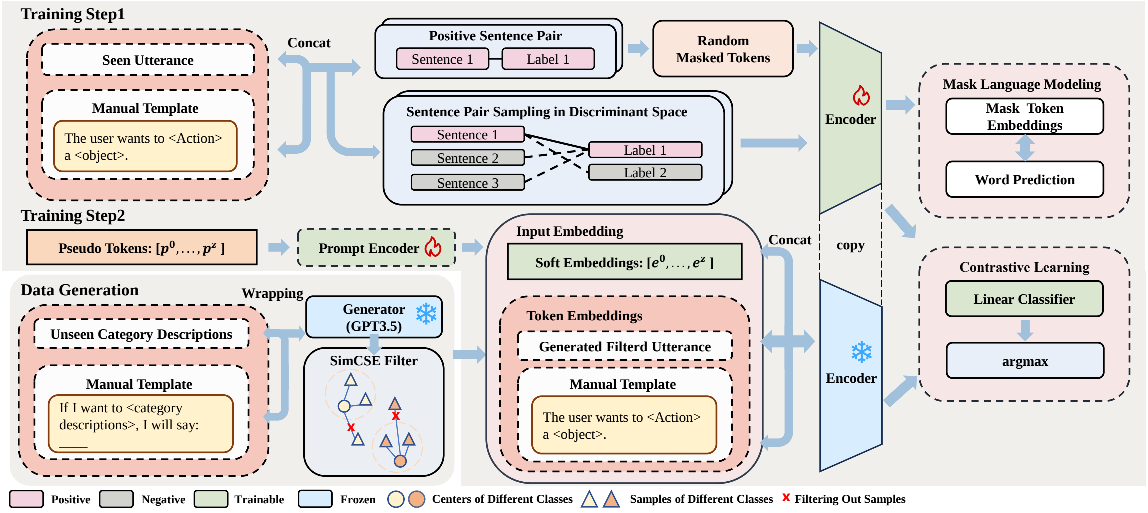

Figure 1: The framework of our pairwise prompt-based tuning model. There are two steps to train this model: base model training and parameter efficient fast adaptation. The first step is to concatenate the utterances and the intent labels with a template, then train the encoder by predicting randomly masked tokens for positive sentence pairs as well as identifying the positive and negative sentence pairs with a hybrid contrastive learning loss. The second step is to generate unseen data by providing category descriptions to GPT3.5 and filter these generated data with SimCSE, then fine-tune the prompt encoder by adding some soft prompts to adapt to incoming unseen intents.

ods like ZeroGen (Ye et al., 2022) and PromptMix (Sahu et al., 2023) assume that there are no seen data available, thus they implement data augmentation, which leverages large language models and label descriptions to generate synthetic data, then train a supervised model to identify unseen data.

2.2 Contrastive Learning

The effectiveness of contrastive learning is prominently demonstrated in both computer vision (Bachman et al., 2019; He et al., 2020; Zhang et al., 2021a; Han et al., 2021) and natural language processing (Gunel et al., 2021; Gao et al., 2021; Yan et al., 2021). This approach typically includes categorizing samples into positive and negative pairs, enabling the model to better capture the data representations. Some studies employ contrastive learning in the prototype space, such as the work of intent recognition (Zhang et al., 2021b). There are also instances of using contrastive learning in the discriminative space to construct positive and negative examples (Sun et al., 2022; Lamanov et al., 2022). Research in (Zhang et al., 2021b) indicates that self-supervised contrastive pre-training and supervised contrastive fine-tuning are highly beneficial for improving intent recognition with limited samples. Simultaneously, the straightforward unsupervised contrastive learning proposed by SimCSE (Gao et al., 2021) has achieved significant success in text alignment tasks.

2.3 Parameter-Efficient Tuning

Fine-tuning large pre-trained language models for downstream tasks has proven effective, but it is also computationally expensive (Pourpanah et al., 2023; Xu et al., 2023; Ding et al., 2023). Recent approaches aim to make transfer learning more parameter-efficient by tuning only a small subset of parameters, which helps to reduce both memory and computational costs while preserving strong performance. Common techniques include: Adapters (Houlsby et al., 2019), which add lightweight layers to the model, enabling task-specific adjustments without the need for full network retraining. LoRA (Low-Rank Adaptation) (Hu et al., 2021), which reduces the number of trainable parameters by decomposing weight matrices into low-rank matrices. P-Tuning (Liu et al., 2021), which freezes the original model parameters and instead employs a small set of trainable continuous prompt embeddings, concatenated with discrete prompts, to guide the model training.

3 The Proposed Method

3.1 Problem Formulation

The goal of generalized zero-shot intent detection is to train a model with utterances of seen intents to recognize unseen intents. Given a set of seen classes $Y^s = \{y^1, ..., y^k\}$ and a set of unseen classes $Y^u = \{y^{k+1}, ..., y^n\}$, where $Y^s \cap Y^u = \emptyset$.

	Intent	Template
VERB + NOUN	book_hotel	the user wants to book a hotel
NOUN	flight_status	the user wants to get a flight status.

Table 1: Illustration of converting intent labels into sentences.

During the training stage, the model can only leverage the samples $D^s = \{(x_i, y_i)\}_{i=1}^{m}$ belonging to the seen classes, where x_i is an utterance and $y_i \in Y^s$ is its corresponding label. Whereas during the test stage, the trained model can be used to predict the samples in both seen and unseen classes.

3.2 Framework Overview

We propose to detect the intent behind each utterance by designing a pairwise prompt-based tuning model, which is trained in two steps. The framework is shown in Figure 1. In step 1, we propose a hybrid contrastive learning loss and a masked language modeling loss to train the base model at both the sentence and token levels. In step 2, we devise a data generation and filtering strategy to produce a handful of unseen data, and use them to fine-tune the base model in a parameter efficient way.

3.3 Pairwise Prompt-Based Tuning Model

We leverage the powerful pre-trained language model to conduct prompt-based tuning with the samples $D^s = \{(x_i, y_i)\}_{i=1}^{m}$ belonging to the seen classes. Considering that the number of classes in the training and test stages are usually inconsistent, to strengthen the generalization ability of the model, we transform the multi-class classification task into a binary classification task which predicts whether there is a contextual relationship between two sentences.

To better align with the NSP task, we expand the intent label y into a coherent sentence, referred to as $template(y)$. This expansion is illustrated in Table 1, as intents typically follow the form of gerunds or noun phrases.

Given an utterance x and an intent label y, we first combine them to obtain a synthetic sentence:

$$T_{(x,y)} = [cls]\ x\ [sep]\ template(y). \tag{1}$$

Then we input the whole sentence $T_{(x,y)}$ into the pre-trained language model Encoder$(\cdot)$ and obtain the $[cls]$ embedding Encoder$(T_{(x,y)})$. Finally, we construct a linear classification module upon the $[cls]$ embedding to obtain the correlation between the utterance x and the intent label y:

$$v(T_{(x,y)}) = \text{Sigmoid}(\boldsymbol{W}\text{Encoder}(T_{(x,y)}) + b), \tag{2}$$

where $v(x, y)$ is a value between 0 and 1, and $\boldsymbol{W}$ and b are model parameters.

3.4 Step 1: Base Model Training

Hybrid Contrastive Learning Loss in Discriminant Space The goal of contrastive learning is to construct positive and negative samples to train the model, improving its representation and discriminative capabilities. Several zero-shot intent detection methods utilize contrastive learning in discriminant space (Sun et al., 2022; Lamanov et al., 2022). However, they typically treat each utterance or intent as an anchor and select a few intents or utterances as negative samples. This approach, due to the vast number of potential negative samples, often fails to guarantee that each intent or utterance has sufficient hard negative samples.

In this section, we define the similarity and dissimilarity (i.e., positive and negative pairs) of samples in discriminant space, which promote the discriminability and robustness of data representations.

Given a positive pair of utterance and intent (x_i^+, y_i^+), we use hybrid negative sampling strategies to sample negative pairs, which involves two parts: 1) sampling a set of hard negative utterances for intent y_i^+: $N_i^{uttr} = \{(x_j^-, y_i^+)\}_{j=1}^{k_1}$; 2) sampling a set of hard negative intents in terms of utterance x_i^+: $N_i^{intent} = \{(x_i^+, y_j^-)\}_{j=1}^{k_1}$.

In the first scenario, for each positive pair (x_i^+, y_i^+), we use a pre-trained sentence similarity learning model SimCSE-RoBERTa-large (Gao et al., 2021) to compute the cosine similarity between each utterance and y_i^+. Then, we select the top 100 out-of-class utterances with highest cosine similarity values with y_i^+ , and randomly choose k_1 samples from the top 100 out-of-class utterances to construct k_1 negative pairs $N_i^{uttr} = \{(x_j^-, y_i^+)\}_{j=1}^{k_1}$.

In the second scenario, we adopt a similar strategy to sample hard negative intents y_j^- for a given utterance x_i^+. In particular, we use SimCSE-RoBERTa-large to calculate the similarity score between each intent and x_i^+, then select the top k_1 intents besides y_i^+ that have higher similarity values. Finally we obtain k_1 negative pairs $N_i^{intent} = \{(x_i^+, y_j^-)\}_{j=1}^{k_1}$.

We perform contrastive learning on each positive pair and its corresponding negative pairs and minimize the following loss:

$$\begin{aligned} L^i_{dis} &= -log(v(x^+_i, y^+_i)) \\ &- \sum_{(x^-_j, y^+_i) \in N^{uttr}_i} log(1 - v(x^-_j, y^+_i)) \\ &- \sum_{(x^+_i, y^-_j) \in N^{intent}_i} log(1 - v(x^+_i, y^-_j)). \end{aligned} \tag{3}$$

Masked Language Modeling The hybrid contrastive learning loss in discriminant space enhances the model's ability to distinguish hard negative samples. We aim to further capture token-level connections within positive sample pairs by deriving token representations through contextual understanding.

For each positive utterance-intent pair (x^+_i, y^+_i), we concatenate them with a template using Eq. (1) to obtain the sentence $p_i = T_{(x^+_i, y^+_i)}$. Then we employ the MLM task to randomly mask the sentences of all the positive pairs D^s. Our goal is to predict the masked tokens in each sentence p_i, and minimize the cross-entropy loss between the predicted tokens and true tokens. Here we use L^i_{mlm} as the average cross-entropy loss over all masked tokens for each utterance x_i:

$$L^i_{mlm} = -\frac{1}{|M(p_i)|} \sum_{w_j \in M(p_i)} logP(w_j | \tilde{p_i}), \tag{4}$$

where $\tilde{p_i}$ represents the masked version of the sentence p_i. $M(p_i)$ represents the set of masked tokens in the sentence p_i, and its cardinality is denoted as $|M(p_i)|$.

Training loss In the training phase, for each utterance $x_i \in D^s$, we combine its hybrid contrastive learning loss and masked language modeling loss, and obtain the overall loss function for all the training utterances.

$$\min L = \sum_{x_i \in D^s} L^i_{dis} + \lambda L^i_{mlm}, \tag{5}$$

where $\lambda \in (0, 1]$ is a trade-off hyperparameter.

3.5 Step 2: Parameter Efficient Fast Adaptation

Data Generation via Category Descriptions Large language models have demonstrated outstanding capabilities in generating high-quality text samples (Schick and Schütze, 2021). In this section, we utilize GPT3.5 as a generator to produce unseen data.

Given an unseen category $y^u \in Y^u$, we slightly extend the intent label to obtain a natural sentence l^u, then design a data construction template $T_{gen}(l^u)$ to encapsulate the category description l^u so as to guide the generation of unseen data. Specifically, the template $T_{gen}(l^u)$ takes the form: "If I want to <category description>, I will say :" (Liu et al., 2024). To ensure that the generated tokens end at a logical position, we adopt the "quote-ending" strategy (Schick and Schütze, 2021) to generate m' training data $D^u_{gen} = \{(x^u_i, y^u_i)\}^{m'}_{i=1}$ for each unseen category y^u.

To avoid introducing noise into the generated data, we further use SimCSE-RoBERTa-large to select data x^u_i whose cosine similarity to their corresponding textual intent $template(y^u_i)$ is higher than the threshold ϵ, and obtain the filtered generated dataset:

$$D^u_{gen'} = \{(x^u_i, y^u_i) \mid sim(x^u_i, template(y^u_i)) > \epsilon\}, \tag{6}$$

where $template(y^u_i)$ is a sentence converted from the intent label y^u_i through the template in Table 1.

Additionally, to prevent overfitting to unknown categories, we randomly select k' samples for each seen intent category from the initial training set D^s to form $D^s_{k'}$, then combine them with the generated unseen data $D^u_{gen'}$ to constitute the training dataset in step two.

P-tuning After generating a few number of unseen data, we adopt a popular parameter-efficient fine-tuning strategy called P-tuning (Liu et al., 2021) to achieve fast adaptation. In specific, we fix the parameters of the base model and only fine-tune a small number of new parameters to quickly adapt the model to unseen intents.

Firstly, we construct pseudo tokens for soft prompts: $[p_0, \cdots, p_z]$. For each sentence constructed by a template $T_{(x,y)}$, we encapsulate it using the soft prompts as:

$$T^p_{(x,y)} = \{[p_{0:i}],\ T_{(x,y)},\ [p_{i+1:z}]\}. \tag{7}$$

Secondly, we concatenate the embeddings of the original input sequence $T_{(x,y)}$ with z trainable embedding vectors from the prompt encoder, formulating the input as:

$$\{\boldsymbol{e}_0, \cdots, \boldsymbol{e}_i,\ enc(T_{(x,y)}), \boldsymbol{e}_{i+1}, \cdots, \boldsymbol{e}_z\}, \tag{8}$$

Dataset	#Classes				#Samples		Average Length	Balanced
	seen	unseen	overall	unseen%	total	average		
Atis	12	5	17	30%	4972	245	11.44	False
MultiWoZ	8	3	11	30%	27449	2495	11.07	False
Clinc150	112	38	150	25%	22500	150	8.31	True
Banking77	57	20	77	25%	13083	170	11.91	True

Table 2: Dataset statistics. "Unseen%" indicates the percentage of unseen intents in the total classes. The "average #Samples" indicates the average number of samples per class. "Balanced" indicates whether the dataset is balanced or imbalanced.

where $\boldsymbol{e}_i$ is the embedding of the pseudo token p_i obtained through the prompt encoder, and $enc(T_{(x,y)})$ is the input embeddings of the sentence $T_{(x,y)}$.

For the prompt encoder, we choose a bidirectional Long Short-Term Memory (LSTM) network paired with a two-layer Multilayer Perceptron (MLP) with ReLU activation.

$$\begin{aligned} \boldsymbol{e}_i &= \text{MLP}([\overrightarrow{\boldsymbol{e}_i} \,||\, \overleftarrow{\boldsymbol{e}_i}]) \\ &= \text{MLP}([\text{LSTM}(\boldsymbol{e}_{0:i})||\text{LSTM}(\boldsymbol{e}_{i+1:z})]), \end{aligned} \tag{9}$$

where $||$ denotes the concatenation operator.

We employ the same hybrid negative sampling strategy as in section 3.4 to construct positive and negative examples for contrastive learning, updating only the prompt encoder's parameters. Using a small amount of generated unseen data and freezing the base model, we can fine-tune and save a minimal number of parameters, enabling fast adaptation to new intents.

3.6 Inference

In the test phase, we input an utterance x^{test} combined with the trained soft prompts into the model, and calculate its correlation value with each intent in both seen and unseen class sets, then predict its label by selecting the intent which yields the highest correlation score.

$$y^* = \underset{y_j \in Y}{\arg\max}\; v(T^p_{(x^{test}, y_j)}), \tag{10}$$

where $Y = Y^s \cup Y^u$ is a class set that involves all the intents.

4 Experiment

4.1 Dataset

We conduct experiments across four widely-used English intent recognition datasets. The statistics and data splitting of all the datasets are presented in Table 2.

Atis (Hemphill et al., 1990): is a classic dataset for natural language processing (NLP) and dialogue system research. Atis encompasses dialogues related to air travel. It covers 17 fine-grained intents involving flight information, ticket reservations, and airline details. It is noteworthy that the number of utterances of each intent in Atis is highly imbalanced. Specifically, the "flight" category constitutes approximately 73.73% of the whole Atis dataset.

MultiWoZ (Budzianowski et al., 2018): is a well-known and publicly available dataset. We use the recent version 2.2 of MultiWoZ in our experiments, which contains utterances with 11 intents. Following the previous works (Siddique et al., 2021; Lamanov et al., 2022), we keep the utterances that have intents expressed by users.

Clinc150 (Larson et al., 2019): is a recently published intent detection dataset that includes 22,500 in-scope queries covering 150 intent classes from 10 domains.

Banking77 (Casanueva et al., 2020): is a fine-grained intent detection dataset with 77 intents, which is collected from banking dialogues comprising 13,083 utterances.

4.2 Data Splitting

Data Splitting for Classes. We adopt the partitioning approach proposed by (Zhang et al., 2022) for the Atis dataset. For the MultiWoZ, Clinc150, and Banking77 datasets, we follow the approach outlined in (Lamanov et al., 2022). Specifically, for Atis and MultiWoZ, we choose 30% of intents as unseen intents (5 out of 17 and 3 out of 11). For CLINC and BANKING, we select 25% of intents as unseen intents (38 out of 150 and 20 out of 77). It is noteworthy that unlike conducting 10 tests under a fixed splitting of classes, we implement 10 tests by randomly splitting the classes, i.e., we set the seed value from 0 to 9 to split the seen and unseen classes, ensuring a random and diverse

Method	**Atis**					
	unseen		seen		overall	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
ZSDNN+CTIR	5.76±6.73	6.27±8.47	69.42±24.80	72.43±23.79	61.38±27.06	51.60±25.10
CapsNet+CTIR	5.85±7.45	9.72±12.10	95.12±1.78	**96.82±1.34**	62.26±38.93	59.77±40.01
RIDE+PU	48.29±20.48	53.48±20.70	86.40±7.65	91.92±4.99	71.95±20.44	74.04±15.88
LTA	34.92±9.10	46.74±10.64	**97.02±1.46**	71.64±25.96	70.00±20.24	72.57±14.31
SP RoBerta+template	39.70±21.30	48.80±21.98	94.94±7.21	96.23±4.35	81.28±10.25	81.23±8.38
AGCR	39.43±20.92	43.46±24.12	67.40±11.81	76.85±19.11	67.37±12.80	62.42±13.79
Ours (step1)	43.23±21.90	52.45±22.78	96.23±5.18	96.40±5.08	82.71±10.49	82.20±8.65
Ours (step1+step2)	**54.55±21.41**	**62.78±21.90**	95.58±5.04	96.01±4.70	**85.54±7.85**	**85.54±6.86**

Table 3: Results on Atis.

Method	**MultiWoZ**					
	unseen		seen		overall	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
ZSDNN+CTIR	32.29±9.22	45.80±12.99	86.30±0.83	90.21±0.92	72.14±4.93	68.80±6.72
CapsNet+CTIR	0.18±0.07	0.36±0.14	93.50±0.95	94.12±0.84	68.73±5.05	60.36±6.81
RIDE+PU	74.28±9.61	79.46±13.72	89.84±6.85	91.43±7.36	81.63±3.93	82.03±6.35
LTA	69.84±12.87	76.90±12.81	91.55±2.75	82.35±4.96	79.08±9.93	78.96±10.00
SP RoBerta+template	62.40±23.10	72.20±17.50	**94.10±1.10**	**94.80±1.00**	78.50±10.30	78.20±10.50
AGCR	75.36±9.41	83.45±7.83	84.37±6.98	85.74±5.24	80.28±6.91	79.68±7.46
Ours (step1)	72.86±8.21	81.03±6.64	87.71±5.12	89.45±4.50	80.52±4.53	81.07±4.37
Ours (step1+step2)	**75.43±7.56**	**83.61±6.47**	91.70±1.53	92.86±1.04	**82.60±4.47**	**82.99±4.49**

Table 4: Results on MultiWoz.

Method	**Clinc150**					
	unseen		seen		overall	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
ZSDNN+CTIR	29.84±3.04	38.81±3.60	78.48±1.10	79.12±0.93	66.21±1.04	63.22±1.31
CapsNet+CTIR	4.25±2.62	7.15±2.03	92.68±0.90	93.24±0.95	69.86±0.82	64.13±0.50
RIDE+PU	51.69±24.58	50.69±25.37	53.59±28.09	54.82±26.57	52.69±25.89	51.94±26.30
LTA	59.49±2.86	67.09±2.68	93.89±1.09	82.56±1.87	75.64±1.93	74.35±2.16
SP RoBerta+template	69.20±3.10	76.60±2.80	92.70±0.90	93.10±0.80	80.20±1.80	81.70±1.50
AGCR	62.08±3.16	62.19±2.84	77.35±0.92	78.24±0.76	70.58±1.93	71.63±1.56
Ours (step1)	71.68±3.38	77.83±3.42	95.55±0.87	96.11±0.87	82.88±1.76	82.14±2.01
Ours (step1+step2)	**73.23±2.69**	**79.13±2.77**	**95.74±0.91**	**96.33±0.78**	**85.05±0.93**	**84.16±1.14**

Table 5: Results on Clinc150.

Method	**Banking77**					
	unseen		seen		overall	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
ZSDNN+CTIR	11.87±1.83	29.10±1.47	39.31±1.05	41.36±0.74	31.45±1.10	29.14±1.17
CapsNet+CTIR	1.36±0.42	3.19±0.80	79.67±1.52	80.56±1.22	58.89±1.72	52.63±1.34
RIDE+PU	31.37±14.59	30.16±16.12	21.16±10.04	23.43±9.82	26.17±11.75	25.12±12.73
LTA	56.02±1.97	64.89±2.31	**89.46±0.76**	76.59±1.93	71.32±1.39	70.25±1.85
SP RoBerta+template	58.38±3.76	67.89±3.09	88.61±0.99	**90.20±0.78**	72.22±1.97	71.70±2.08
AGCR	49.65±3.32	56.37±2.22	73.64±1.77	78.42±1.17	63.28±2.15	62.55±2.31
Ours (step1)	62.64±3.65	70.23±3.67	83.39±1.60	85.97±1.33	72.26±1.63	71.78±1.83
Ours (step1+step2)	**62.82±3.34**	**70.48±3.42**	86.11±1.22	88.49±0.77	**73.40±1.52**	**72.79±1.65**

Table 6: Results on Banking77.

evaluation of our proposed model and baselines.

Data Splitting for Training and Test Sets. We randomly select 70% of samples for each seen intent. 90% of them are assigned to the training set, and the remaining 10% of samples constitute the validation set. The test set comprises the remaining 30% of samples in each seen class and all the samples in unseen classes.

4.3 Baselines and Evaluation Metrics

We compare our method with existing zero-shot intent detection methods including ZSDNN (Kumar et al., 2017), CapsNet (Xia et al., 2018), CTIR (Si et al., 2021), RIDE+PU (Siddique et al., 2021; Su et al., 2021), LTA (Zhang et al., 2022), SP RoBerta+template (Lamanov et al., 2022), and AGCR (Liu et al., 2024). To enhance the performance of ZSDNN and CapsNet in the GZID setting, we apply CTIR to them, and denote them as ZSDNN+CTIR and CapsNet+CTIR. The details of these baselines are provided in Appendix A.1.

Following the previous GZID methods, we use accuracy (Acc) and weighted F1 score (F1) to evaluate classification performance. Both metrics are computed with the average value weighted by the sample ratio of the corresponding class.

4.4 Implementation Details

Following previous works, we conduct 10 tests using seeds ranging from 0 to 9 to randomly split the seen and unseen classes, dividing the data into training, validation, and test sets. We then report the mean results and standard deviations for seen, unseen, and overall intents separately. All experiments are performed on NVIDIA RTX A100 GPU with 80GB VRAM. More details about the hyperparameter setting are provided in Appendix A.2.

4.5 Result Analysis

The results on four intent detection datasets are presented in Table 3, Table 4, Table 5, and Table 6, respectively. The highest results are highlighted in bold, and the second-highest results are underlined. From the results, the following observations can be made.

(1) Our model exhibits remarkable performance on both balanced and imbalanced datasets without overfitting to the seen intents, demonstrating its robustness in handling the GZID tasks. Compared to the state-of-the-art baseline SP Roberta+template, the average accuracy and F1 score on overall intents are increased by 3.60% and 3.16%, respectively.

(2) Training on step 1 has already surpassed the baselines, although the network architecture of the base model is similar with SP Roberta+template, we have further boosted the model performance via a combination of hybrid contrastive learning and masked language modeling.

(3) Fine-tuning the prompt encoder in step 2 adjusts a few parameters on the generated unseen data, further improving the ability of detecting unseen intents without compromising the effectiveness in detecting seen intents.

(4) ZSDNN+CTIR and CapsNet+CTIR demonstrate competitive results on seen intents. However, under the setting of 10 random data splitting, the average performance on unseen intents is relatively lower. This indicates that their models may overfit to the seen intents, resulting in a decrease in the overall performance.

(5) RIDE+PU, by integrating a robust external knowledge source in the form of a knowledge graph, mitigates model dependency and exhibits satisfactory results even under unseen intents. However, the standard deviations on the four datasets indicate that the effectiveness of RIDE+PU heavily relies on the data splitting.

(6) LTA and AGCR directly incorporate virtual unseen categories or generate unseen data during training, yet their performance remains inferior to our model (step1), further confirming the generalization capability of our model.

4.6 Ablation Study

To verify the effectiveness of hybrid contrastive learning and masked language modeling in the first step, we conduct ablation study using seeds from 0 to 9 to randomly split seen and unseen classes with average results shown in Table 7 and Table 8. The highest results are in bold, and the second-highest results are underlined. "w/o mlm" indicates the removal of the masked language modeling loss, while "w/o hard intent" and "w/o hard utter" represent using only hard utterances or hard intents as negative samples in the hybrid contrastive learning loss. To match the number of negative samples in the original version, the number of hard utterances in "w/o hard intent" is set to $2k_1$. In "w/o hard utter", the number of negative intents is set to $k_1' = min(N_{seen} - 1, 2k_1)$, as the number of seen intents N_{seen} may be smaller than $2k_1$.

The results in Table 7 and Table 8 show a signifi-

Configuration	Atis						MultiWoZ					
	unseen		seen		overall		unseen		seen		overall	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Ours (step1)	43.23	52.45	96.23	96.40	82.71	82.20	72.86	81.03	87.71	89.45	80.52	81.07
w/o mlm	37.21	45.68	97.07	97.61	**82.77**	81.62	71.12	80.02	**90.92**	**92.01**	79.77	80.43
w/o hard intent	**43.42**	**53.63**	**97.12**	**97.72**	82.57	**82.54**	72.71	81.15	90.24	91.62	80.50	81.07
w/o hard utter	34.84	43.60	94.40	94.10	77.91	77.29	**73.10**	**81.49**	90.82	91.07	**81.35**	**81.88**

Table 7: Ablation study on Atis and MultiWoZ.

Configuration	Clinc150						Banking77					
	unseen		seen		overall		unseen		seen		overall	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Ours (step1)	**71.68**	**77.83**	**95.55**	**96.11**	**82.88**	**82.14**	**62.64**	**70.23**	**83.39**	**85.97**	**72.26**	**71.78**
w/o mlm	70.85	77.05	93.47	94.37	81.47	80.73	61.84	69.85	83.31	85.77	71.67	71.26
w/o hard intent	70.76	76.88	92.52	93.45	80.97	80.32	57.85	65.88	75.42	78.22	65.89	65.85
w/o hard utter	69.22	75.38	88.75	89.89	78.38	77.76	59.59	67.52	73.99	76.19	66.18	65.99

Table 8: Ablation study on Clinc150 and Banking77.

cant drop in performance for recognizing unseen intents when a submodule is removed. (1) Removing the masked language modeling loss (i.e., w/o mlm) reveals that this task enhances the model's ability to classify unseen intents, showing that incorporating a cloze task within positive sample pairs at token level improves generalization in the GZID setting. (2) The performance of only utilizing hard utterances or hard intents as negative samples (i.e., w/o hard intent or w/o hard utter) varies depending on the datasets, with some datasets benefiting more from hard utterance sampling and some benefiting more from hard intent sampling. Nevertheless, the overall performance across the four datasets indicates that a hybrid negative sampling strategy by combining both hard utterances and hard intents is more useful for discriminating the unseen intents.

5 Conclusion

In this paper, we explore intent detection in dynamic development scenarios with continuously evolving novel intents. To address this challenge, we propose a pairwise prompt-based tuning model with two steps: base model training and parameter efficient fast adaptation. The first step uses hybrid contrastive learning with informative negative samples, while enhancing token-level representation through masked language modeling. The second step generates and refines unseen data, utilizing parameter-efficient fine-tuning for rapid adaptation to new intents. Extensive experiments on four intent detection datasets validate the superiority of our method in the GZID setting.

Limitations

Firstly, our approach relies on templates to convert intent labels into sentences. We only use a simple template here, introducing more precise templates or designing specialized templates for specific domains may further enhance the results. Secondly, in binary classification tasks, as the number of intents increases, the computational resources also escalate. Maintaining a balance between classification performance and resource consumption is worth further exploration in the future.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62206038, 62106035), Liaoning Binhai Laboratory Project (No. LBLF-2023-01), Chunhui Project Foundation of the Education Department of China (No. HZKY20220419), and Xiaomi Young Talents Program.

References

Yuval Atzmon and Gal Chechik. 2019. Adaptive confidence smoothing for generalized zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11671–11680. IEEE.

Philip Bachman, R. Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Advances in*

Neural Information Processing Systems (NeurIPS), pages 15509–15519.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 93–104. ACM.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5016–5026. Association for Computational Linguistics.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. *CoRR*, abs/2003.04807.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 830–835. AAAI Press.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mac. Intell.*, 5(3):220–235.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. 2021. Contrastive embedding for generalized zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2371–2381. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735. Computer Vision Foundation / IEEE.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Proceedings of Speech and Natural Language Workshop*. Morgan Kaufmann.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Anjishnu Kumar, Pavankumar Reddy Muddireddy, Markus Dreyer, and Björn Hoffmeister. 2017. Zero-shot learning across heterogeneous overlapping domains. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2914–2918. ISCA.

Dmitry Lamanov, Pavel Burnyshev, Ekaterina Artemova, Valentin Malykh, Andrey Bout, and Irina Piontkovskaya. 2022. Template-based approach to zero-shot intent recognition. *CoRR*, abs/2206.10914.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316. Association for Computational Linguistics.

Han Liu, Xiaotong Zhang, Lu Fan, Xuandi Fu, Qimai Li, Xiao-Ming Wu, and Albert Y. S. Lam. 2019. Reconstructing capsule networks for zero-shot intent classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4798–4808. Association for Computational Linguistics.

Han Liu, Siyang Zhao, Xiaotong Zhang, Feng Zhang, Wei Wang, Fenglong Ma, Hongyang Chen, Hong Yu, and Xianchao Zhang. 2024. Liberating seen classes: Boosting few-shot and zero-shot text classification via anchor generation and classification reframing. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18644–18652. AAAI Press.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *CoRR*, abs/2103.10385.

Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and Q. M. Jonathan Wu. 2023. A review of generalized zero-shot learning methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4051–4070.

Gaurav Sahu, Olga Vechtomova, Dzmitry Bahdanau, and Issam H. Laradji. 2023. Promptmix: A class boundary augmentation method for large language model distillation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5316–5327. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6943–6951. Association for Computational Linguistics.

Olga Seminck. 2023. Conversational AI: dialogue systems, conversational agents, and chatbots by michael mctear. *Comput. Linguistics*, 49(1):257–259.

Qingyi Si, Yuanxin Liu, Peng Fu, Zheng Lin, Jiangnan Li, and Weiping Wang. 2021. Learning class-transductive intent representations for zero-shot intent detection. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3922–3928. ijcai.org.

A. B. Siddique, Fuad T. Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized zero-shot intent detection via commonsense knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1925–1929. ACM.

Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 935–943.

Guangxin Su, Weitong Chen, and Miao Xu. 2021. Positive-unlabeled learning from imbalanced data. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2995–3001. ijcai.org.

Yi Sun, Yu Zheng, Chao Hao, and Hangping Qiu. 2022. NSP-BERT: A prompt-based few-shot learner through an original pre-training task - - next sentence prediction. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 3233–3250. International Committee on Computational Linguistics.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2977–2992. Association for Computational Linguistics.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3090–3099. Association for Computational Linguistics.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. 2019. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2251–2265.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *CoRR*, abs/2312.12148.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 5065–5075. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11653–11669. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3912–3921. Association for Computational Linguistics.

Haode Zhang, Yuwei Zhang, Li-Ming Zhan, Jiaxin Chen, Guangyuan Shi, Xiao-Ming Wu, and Albert Y. S. Lam. 2021a. Effectiveness of pre-training for few-shot intent classification. In *Findings of the Association for Computational Linguistics: (EMNLP)*, pages 1114–1120. Association for Computational Linguistics.

Jian-Guo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip S. Yu. 2021b. Few-shot intent detection via contrastive pre-training and fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1906–1912. Association for Computational Linguistics.

Yiwen Zhang, Caixia Yuan, Xiaojie Wang, Ziwei Bai, and Yongbin Liu. 2022. Learn to adapt for generalized zero-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 517–527.

A Appendix

A.1 Baseline Description

ZSDNN (Kumar et al., 2017) employs squared Euclidean distance and triplet loss to maintain margins between different classes, with the label embedding as the anchor and the closest sample as the negative in each triplet.

CapsNet (Xia et al., 2018) leverages capsule neural networks for effective adaptation to new intents via knowledge transfer. It extends capsule networks to text modeling, using a hierarchical approach to extract and aggregate semantic information from utterances.

CTIR (Si et al., 2021) is a class-transductive framework that incorporates unseen label names during training, expanding the prediction space to involve unseen classes, which can be integrated with existing zero-shot intent detection methods.

RIDE+PU (Siddique et al., 2021; Su et al., 2021) leverages commonsense knowledge from ConceptNet and Positive-Unlabeled learning to capture deep semantic relationships between utterances and intent labels.

LTA (Zhang et al., 2022) avoids overfitting on seen classes by training an adaptive classifier using both seen and virtual unseen classes to simulate a generalized zero-shot learning scenario.

SP RoBerta+template (Lamanov et al., 2022) extends labels into sentences transforming to a binary classification task and uses contrastive learning to enhance its discriminative capability.

AGCR (Liu et al., 2024) employs a generative model to produce pseudo samples for unseen categories, selecting typical ones as category anchors to avoid negative transfer and capture the essence of unseen categories through their descriptions.

A.2 Hyperparameter Setting

For the baselines, we conduct experiments by utilizing the open-source code. We follow their experimental setting in the original paper and carefully adjust the hyperparameters based on the validation set. For SP Roberta+template, four different templates are tested in their experiments, and we choose the $d1\_template$ that exhibits the best performance.

For our model, the hyperparameters are set according to the performance of validation set. In the first step, for the linear classification module, we employ a dropout layer with the dropout rate of 0.5, a linear layer, and a Sigmoid function to map the results between 0 and 1. In discriminative space, the number of negative samples in terms of hard utterances and hard intents is set to $k_1 = 7$. For the whole pairwise prompt-based tuning model, the learning rates are configured separately for Atis, MultiWoZ, Clinc150, and Banking77 as [2e-5, 2e-6, 2e-5, 5e-5], and the batch size is set to [16,16,32,32]. We set the hyperparameters λ as 0.3. RoBERTa-base is chosen as our pre-trained language model to align with the state-of-the-art method (Lamanov et al., 2022). The dimension of the embeddings in the hidden layer is 768. The dropout rate in RoBERTa-base is set to its default value 0.1. In the MLM task of our model, the masking ratio and strategy are hyperparameters. Conventionally, pre-trained models such as BERT and RoBERTa choose a 15% probability for masking tokens of the input, following the 80-10-10 masking allocation strategy: 80% are replaced with $[MASK]$ tokens, 10% are replaced by some random words, and 10% of the words are unchanged. However, based on the investigation of (Wettig et al., 2023), we opt for a 20% probability as the masking ratio instead of 15%, and modify the 80-10-10 masking allocation strategy by 100% replacing the masked tokens with $[MASK]$ by default.

In the second step, the parameters used for invoking the GPT-3.5 API are set to p=0.9 and temperature=1.5. The number of generated data for each unseen intent is set to $m' = 50$, and the number of selected data for each seen intent is set to $k' = 50$ as well. For filtering the generated unseen data and P-tuning, we customize thresholds, prompt lengths, and learning rates for each dataset: Atis (ϵ: 0.5, prompt length: 6, learning rate: 2e-4), MultiWoZ (ϵ: 0.6, prompt length: 4, learning rate: 2e-6), Clinc150 (ϵ: 0.6, prompt length: 4, learning rate: 2e-4), and Banking77 (ϵ: 0.6, prompt length: 6, learning rate: 5e-5).

A.3 Selection of k_1

To analyze the impact of different k_1 values in contrastive learning, we conduct experiments on both the imbalanced MultiWOZ dataset and the balanced Bank77 dataset using seed 0. We evaluate our method by setting $k_1 = 1, 3, 5, 7, 9$. Note that since MultiWOZ only use 8 intent categories for training, k_1 cannot be set to 9. The accuracy results on overall intents are shown in Table 9, which indicates that the best performance is achieved when k_1 is around 7.

Method	1	3	5	7	9
MultiWOZ	85.49	85.62	86.93	88.41	-
Bank77	75.69	78.41	77.43	77.38	76.47

Table 9: Selection of k_1

A.4 Effectiveness of Loss Function

We conduct experiments to compare the Binary Cross-Entropy (BCE) loss with the ranking-based contrastive loss using softmax (RBC). We run both methods by setting seed value from 0 to 9. The average accuracy results on unseen, seen, and overall intents are shown in Table 10, which shows that the BCE loss is comparable with the RBC loss.

Method	BCE Loss	RBC Loss
MultiWOZ	75.43 / 91.70 / 82.60	73.48 / 93.71 / 83.36
Bank77	62.82 / 86.11 / 73.40	57.37 / 85.25 / 71.57

Table 10: Performance comparison of BCE and RBC losses on unseen, seen, and overall intents.