# VSLIM: A Vietnamese Explicit Slot-Intent Mapping for Joint Multi-Intent Detection and Slot Filling

Phong Chung[1,2], Kha Le-Minh[2,3], Xuan-Bach Le[1,2⋆], and Tho Quan[1,2]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
{cdphong.sdh232,lexuanbach,qttho}@hcmut.edu.vn
[2] Vietnam National University Ho Chi Minh City, Linh Xuan Ward, Ho Chi Minh City, Vietnam
[3] University of Information Technology, Ho Chi Minh City, Vietnam
23520664@gm.uit.edu.vn

**Abstract.** Multi-intent detection and slot filling are fundamental tasks of natural language understanding in task-oriented dialog systems. Early approaches treated them as separate tasks, which undermines the direct connection between intents and their associated slots. This limitation becomes more pronounced when multiple intents are expressed within a single utterance. In the Vietnamese language landscape, research on this topic remains limited, largely due to its low-resource status. To address this gap, we introduce VSLIM, a joint model designed for multi-intent detection and slot filling in Vietnamese. Inspired by the SLIM framework [2], VSLIM builds on its foundation with a biaffine classifier that more directly captures the relationship between intents and slots. This design allows the model to better understand and represent the dependencies across sequence labels in multi-intent settings. Experiments on the Vietnamese PhoATIS dataset and our newly introduced VPED corpus show that VSLIM outperforms strong NLU baselines, highlighting its potential for improving Vietnamese task-oriented dialog systems. We publish our VSLIM implementation and VPED at https://github.com/dongphong543/VSLIM.

**Keywords:** Multi-intent detection · Slot filling · Vietnamese language understanding · Joint learning · Explicit mapping.

## 1 Introduction

Task-oriented dialog (ToD) systems rely mainly on natural language understanding (NLU) components to extract structured semantic information from user utterances. Two core tasks of **intent detection (ID)** and **slot filling (SF)** play a foundational role for effective dialog management and response generation. For example, in the sentence "Spend $10 on Apple Music subscription today",

⋆ Corresponding author.

the system would classify the intent as "Add_Expense" and assign BIO tags [18] such as `[O B-price O B-item I-item I-item B-date]`. Early studies approached ID and SF as separated problems, employing models such as convolutional neural networks [9], hierarchical attention mechanisms [26], and sequential CNNs for spoken language understanding [21]. Later, joint frameworks such as Capsule-NLU and JointBERT-CRF demonstrated that sharing contextual representations between the two subtasks significantly improves both performance and robustness [28,3]. However, most existing approaches still rely on the single-intent assumption. In real-life conversations, users often express multiple intents in a single turn of speech, posing challenges for many ToD systems.

Most research in multi-intent understanding can be categorized into two main approaches: *implicit* and *explicit* slot-intent mapping, both with different advantages and limitations. However, the lack of annotated data, especially for explicit mapping, makes this approach difficult to develop and evaluate effectively, limiting its adoption in real world task-oriented dialog systems. This gap is more acute in Vietnamese ToD, where NLU research is limited by scarce annotated data, word segmentation errors, and diacritic inconsistencies. The limited availability of multi-intent datasets further constrains research on explicit slot–intent modeling.

To address these challenges, we introduce **VSLIM** (Vietnamese SLot-Intent Mapping) to jointly handle intent detection and slot filling for Vietnamese language. Drawing inspiration from the SLIM framework [2], VSLIM explicitly models how specific intents relate to their corresponding slots. We utilize the state-of-the-art PhoBERT [13], a pre-trained language model optimized for the Vietnamese language, as the encoder to generate rich contextualized token representations that support accurate multi-intent detection and slot identification. We extend it with a *biaffine classifier*, following [5], to capture fine-grained token-level interactions between intents and slots. This architectural extension allows the system not only to improve accuracy, but also to make intent-slot relationships more interpretable. We assess our model using PhoATIS [4] and a newly curated dataset, which introduces novel measures of slot and intent similarity. To the best of our knowledge, this is the first comprehensive effort to develop such an explicit slot-intent modeling in Vietnamese dialogue systems that support multiple intents.

Our contributions are summarized as follows:

1. We propose **VSLIM**, a unified approach that explicitly captures the relationships between slots and intents to improve Vietnamese task-oriented dialog understanding, particularly in multi-intent settings.
2. We create a new **Vietnamese multi-intent benchmark** by developing Vietnamese Personal Expense Dataset - **VPED**, providing the first large-scale resource for evaluating explicit slot-intent correspondence.
3. We comprehensively evaluated VSLIM against strong baselines, such as AGIF [17], JointIDSF [4], MISCA [15]... Our results demonstrate overall improvements in both single-intent and multi-intent scenarios.

The remainder of this paper is organized as follows. In Section 2, we review existing work on intent detection and slot filling, with a focus on approaches to multi-intent modeling. Section 3 introduces our proposed VSLIM framework. In Section 4, we describe our experimental setup and discuss the results. Section 5 concludes the work and suggests future research directions.

## 2 Related Work

This section reviews three major research areas relevant to this study: ID and SF, multi-intent modeling paradigms, and Vietnamese NLU. Although prior work explores shared representations and structured intent-slot mappings, most focus on single-intent settings. Vietnamese ToD research, in particular, lacks multi-intent datasets and explicit models. VSLIM aims to fill this gap by leveraging PhoBERT and a biaffine linker to enable interpretable multi-intent understanding.

### 2.1 Intent Detection and Slot Filling

Early ToD systems treated ID and SF as independent tasks as they typically trained separate models for each. Intent classification commonly used convolutional neural networks (CNNs) or hierarchical recurrent architectures such as RNNs and BiLSTMs to capture sentence-level semantics [9,26]. Slot labeling, on the other hand, was often handled by CRF-based sequence taggers that operated token by token [18,10,12,21].

Pipeline approaches performed reasonably well on simpler datasets. However, later joint models for ID and SF showed that integrating the two tasks leads to better overall results, as incorporating intent context information into slot filling can boost performance. Joint modeling frameworks were introduced to learn ID and SF simultaneously by leveraging shared semantic representations. Models like Capsule-NLU [28] captured hierarchical relationships between intents and slots using dynamic routing between capsule layers. JointBERT [3] further advanced this approach by utilizing pre-trained BERT embeddings to jointly perform both tasks with a unified encoder. Subsequent models introduced refinements such as CRF decoders [7,23,11] for more accurate slot predictions and graph-based interaction layers [17,29] to strengthen the dependencies between tasks. These joint models significantly improved performance and reduced cascading errors by embedding intent-aware context into slot predictions. However, many of the above models still assume that each utterance conveys only a single intent, or inherit a single-intent-like approach. This assumption often falls short in real-world applications, where a single utterance can express multiple intents at the same time.

### 2.2 Multi-Intent Modeling

Understanding multi-intent utterances is essential for the development of task-oriented dialog systems, as users often convey multiple goals within a single

conversational turn. Research in this domain has evolved largely around two key paradigms: *implicit* and *explicit* mapping approaches. Implicit models usually employ shared attention mechanisms or uniform intent distributions across tokens [17,15,22], allowing efficient joint learning of slot filling and intent detection. However, because these methods do not explicitly model this connection, they fail to provide interpretable mappings between slots and their associated intents. By contrast, explicit models establish clearly defined slot–intent mappings, i.e., they show which slot corresponds to which intent. This additional information makes the results more interpretable, and often boosts performance in complex cases, especially when slot types overlap between intents or when semantic roles are ambiguous [6,2]. However, this approach remains relatively underexplored. The main challenges lie in the lack of annotated data and the difficulty of balancing interpretability with computational efficiency, as this usually increases computational complexity, making such models less efficient for large-scale or real-time dialog applications.

Building on the distinction between implicit and explicit paradigms, recent studies have implemented these ideas through concrete model architectures. Implicit approaches - such as DIET [1], JointIDSF [4] and Co-Guiding Net [25] - jointly learn intent and slot representations using shared encoders. Although these models can effectively capture relationships between tasks, they cannot provide effectively the insights of decision-making process. In contrast, explicit models such as SLIM [2] introduce dedicated mapping layers that directly connect slot and intent representations. This structure enhances interpretability and facilitates advanced capabilities, such as cross-intent slot sharing. While explicit architectures show strong promise for improving interpretability and structured reasoning, they are still far less explored than implicit methods. This gap highlights a viable research opportunity for future work to develop multi-intent dialog systems that are more transparent and reliable.

### 2.3 Vietnamese NLU Studies

Research on Vietnamese ToD and NLU remains limited compared to English. The PhoATIS dataset [4] laid the groundwork for Vietnamese intent detection and slot filling, followed by improvements from models like [19] and [20], which leveraged shared PhoBERT encoders and dual attention.

Despite this progress, most studies remain limited to single-intent-like scenarios, with no exploration of explicit slot-intent modeling or multi-intent understanding. Dataset scarcity and linguistic noise still hinder generalization. To address this gap, we introduce VSLIM - the first Vietnamese model to explicitly model multi-intent semantics. VSLIM combines PhoBERT with a biaffine classifier to connect slots and intents, offering both interpretability and strong performance, with simplicity in architecture.

# 3 Methodology

This section introduces our VSLIM framework for joint multi-intent detection and entity recognition. Section 3.1 provides a formal definition of the intent detection and slot filling task, and Section 3.2 describes the overall model architecture.

## 3.1 Problem Statement

We address the task of **joint multi-intent detection** and **slot filling** in task-oriented dialog systems. A user utterance is usually described as an ordered sequence of tokens:

$$x = (x_1, x_2, \ldots, x_N),$$

where $N$ represents the number of tokens in the utterance. For example, the sentence "Today I took the bus for 50k! Show expense statistics this month." can be expressed as such a sequence.

Let

$$I = \{I_1, I_2, \ldots, I_M\}$$

denote the predefined set of **intents**, such as *Add Expense*, *Delete Expense*, or *Show Statistics*.

Also, let

$$S = \{S_1, S_2, \ldots, S_K\}$$

represent the predefined set of **entity (slot) types**, for instance `DATE` (time), `NAME` (transaction name), and `PRICE` (amount).

The model will jointly perform two subtasks:

1. **Intent Detection:** Identify the set of intents $\mathcal{I}_x \subseteq I$ expressed in the utterance.
2. **Entity Recognition:** Assign a slot label $s_n \in S \cup \{\text{O}\}$ to each token $x_n$, where O denotes tokens that do not belong to any entity span under the standard BIO tagging scheme [18].

Following prior works [17,2], we make two assumptions. First, each utterance expresses at least one intent ($|\mathcal{I}_x| \geq 1$). Second, each entity $s_m \in S$ is associated with at most one intent $I_j \in \mathcal{I}_x$, and an entity may span multiple consecutive tokens $(x_{m1}, \ldots, x_{mj})$ under the standard BIO tagging scheme. This joint formulation encourages consistent modeling of intent-slot interactions and contextual dependencies.

An illustrative example is shown in Figure 1. Here, the utterance expresses two intents, `{STAT, ADD} (abbreviations for stat_expense, add_expense)`, and each token is annotated with both slot and intent labels. Notably, there are two distinct occurrences of the slot type `DATE`, each linked to a different intent. This demonstrates the importance of not only identifying slot types correctly, but also assigning them to the appropriate intent - an aspect often overlooked in single-intent or implicit modeling frameworks.

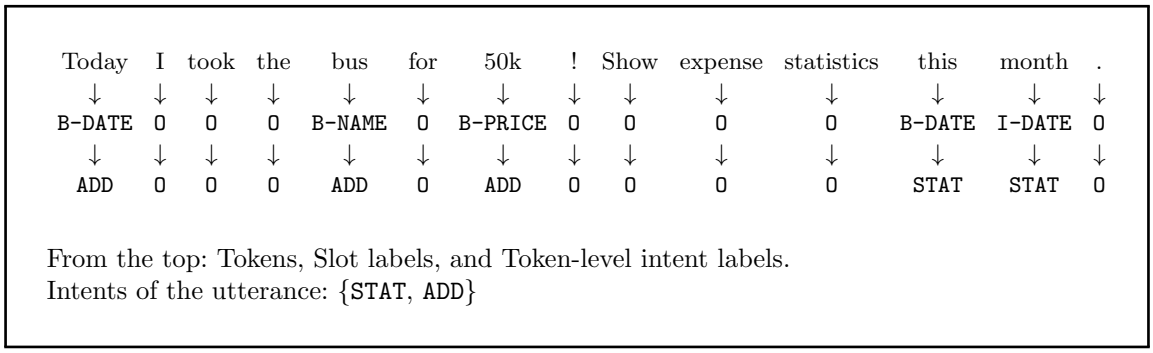


**Fig. 1.** Example of multi-intent detection and slot labeling for an expense-tracking utterance. Each `DATE` slot is associated with a different intent, illustrating the need for explicit slot–intent mapping.

### 3.2 The VSLIM Framework

Building on the SLIM framework proposed by Cai et al. [2], this work presents VSLIM - a novel extension specifically tailored for the understanding of multi-intent Vietnamese dialog. As illustrated in Figure 2, VSLIM extends the original SLIM architecture by incorporating a biaffine classifier. This addition explicitly models the interactions between global intent representations and token-level slot embeddings.

Our system utilizes PhoBERT [13] as the base encoder due to its strong capabilities in Vietnamese language modeling. This pretrained language model produces contextual embeddings that serve as the input for both intent recognition and slot filling. Each user utterance $x$ is concatenated with a `<s>` token (equivalent to `[CLS]` token in BERT) representing sentence-level classification at the beginning and a `</s>` token (equivalent to `[SEP]` token in BERT) indicating sentence termination at the end. It is then tokenized and passed through PhoBERT to produce contextualized embeddings that capture the semantics of the entire utterance:

$$\textbf{Embedding:} \quad h = (h_{\texttt{<s>}}, h_1, \ldots, h_N, h_{\texttt{</s>}}), \quad h_k \in \mathbb{R}^d$$

Subsequently, the resulting embeddings are fed into the intent and slot classifiers to learn global-level intents and token-level slot labels at the same time:

1. **Intent Classifier.** Intent classification task is usually considered as a multi-label classification problem. To predict these intents, the contextualized representation of the special token $h_{\texttt{<s>}}$ is passed through a fully connected layer followed by a sigmoid activation function:

$$y^i = \text{sigmoid}\big(W^i h_{\texttt{<s>}} + b^i\big) \tag{1}$$

Here, $W^i \in \mathbb{R}^{|I| \times d}$ and $b^i \in \mathbb{R}^{|I|}$ are learnable parameters representing the weight matrix and bias vector, respectively. The output vector $y^i \in \mathbb{R}^{|I|}$

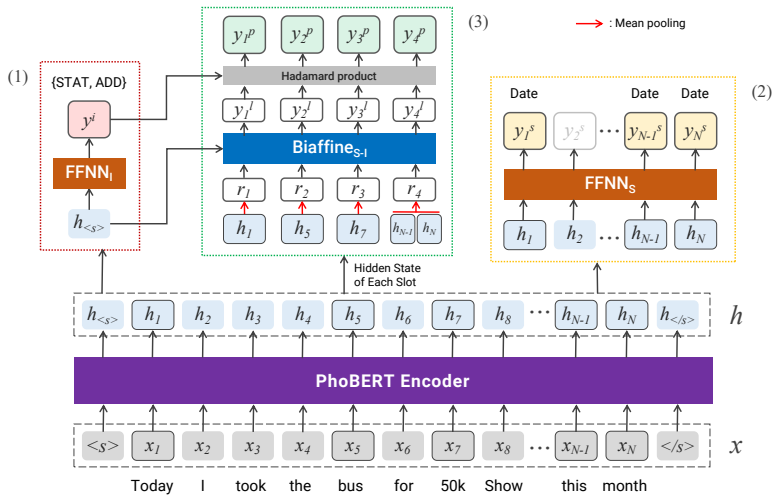


**Fig. 2.** Overview of the VSLIM model with Intent detection (1), Slot filling (2) and Slot-Intent (3) tasks. PhoBERT encodes the input utterance into contextual embeddings. Two FFNNs perform *intent classification* and *slot filling*, while a biaffine classifier models interactions between intent and slot representations.

contains the predicted probabilities for each possible intent label. We apply a threshold of 0.5, which corresponds to an equal probability of the presence or absence of an intent. The model is optimized using the binary cross-entropy loss function $\mathcal{L}_{\text{intent}}$.

2. **Slot Classifier.** Slot filling is often treated as a *sequence labeling* task, where each word from the input is tagged with a slot or entity label. This setup is commonly handled as a token-level multi-class classification problem [12,8,7]. To predict the slot label $y_k^s$ for the $k$-th token, its contextual embedding $h_k$ is passed through a dense layer followed by a softmax function:

$$y_k^s = \text{softmax}(W^s h_k + b^s) \tag{2}$$

With $W^s \in \mathbb{R}^{|S| \times d}$, the resulting vector $y_k^s \in \mathbb{R}^{|S|}$ gives a probability distribution over all possible slot types for token $x_k$. We optimize the model using categorical cross-entropy loss function $\mathcal{L}_{\text{slot}}$.

3. **Token-level Intent Classifier.** To model the relationship between each entity $s_m = \{x_{m_1}, \ldots, x_{m_j}\}$ and the utterance-level intents and perform Slot-Intent mapping task, we first compute its representation via mean pooling:

$$r_m = \frac{1}{j} \sum_{k=1}^{j} h_{m_k} \tag{3}$$

Each pair $(h_{\texttt{<s>}}, r_m)$ is then processed by a biaffine classifier [5], which effectively captures pairwise interactions between global and token-level rep-

**Table 1.** Statistics of PhoATIS and VPED datasets showing intent label counts (*base/compound*), slot label counts and utterance distribution by number of intents.

| Dataset | No. of Intent | No. of Slot | Utterances by no. of Intent (1-5) |
|---|---|---|---|
| PhoATIS | 16 / 24 | 139 | 5805 / 64 / 2 / – / – |
| VPED | 5 / 31 | 16 | 617 / 342 / 294 / 74 / 12 |

resentations [14,27]. The unconstrained token-level intent is computed as:

$$\begin{aligned} y_m^l &= \text{softmax}\left(\text{biaffine}(h_{\texttt{<s>}}, r_m)\right) \\ &= \text{softmax}\left(h_{\texttt{<s>}}^\top \mathbf{W}^b r_m + W^l[h_{\texttt{<s>}} \circ r_m] + b^l\right) \end{aligned} \tag{4}$$

where $\mathbf{W}^b \in \mathbb{R}^{d \times |I| \times d}$, $W^l \in \mathbb{R}^{|I| \times 2d}$, and $b \in \mathbb{R}^{|I|}$. The operator $\circ$ represents vector concatenation. To incorporate sentence-level intent guidance, we apply a Hadamard product (element-wise multiplication) between $y_m^l$ and the global intent prediction $y^i$:

$$y_m^p = y^i \odot y_m^l \tag{5}$$

This process creates an intent distribution for each slot that takes constraints into account. Intents with higher overall confidence are given more weight, allowing us to filter out those that are unlikely. The final constrained token-level intent $y_m^p$ is optimized using the cross-entropy loss $\mathcal{L}_{\text{slot-intent}}$.

*Joint training.* The total loss function is a weighted sum:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{intent}} + \beta \mathcal{L}_{\text{slot}} + \gamma \mathcal{L}_{\text{slot-intent}}, \tag{6}$$

For our setup, we use the weight values $\alpha = 1$, $\beta = 2$, and $\gamma = 1$, which aligns with the scheme outlined by [2]. A higher weight is assigned to the slot-filling loss ($\beta = 2$) due to its increased complexity as a token-level task, and its influence onto both slot-intent and overall frame prediction.

## 4 Experiments and Discussion

In this section, we present a comprehensive experimental analysis of the proposed VSLIM framework. We first describe the datasets (Section 4.1) and experimental setup (Section 4.2). Subsequently, we benchmark VSLIM against several strong baselines on both single- and multi-intent tasks (Section 4.3) and analyze its behavior through detailed ablation studies (Section 4.4).

### 4.1 Dataset

Table 1 summarizes the key statistics of the two datasets used in our experiments: PhoATIS [4] and the Vietnamese Personal Expense Dataset (VPED).

PhoATIS is a Vietnamese adaptation of the ATIS corpus [16]. The dataset contains 4,478 utterances for training, 500 for validation, and 893 for testing. It includes 16 base intents (24 when compound intents are considered) and 139 slot labels [4]. Most of the utterances are single-intent (5,805 out of 5,871, $\approx 99\%$).

**Table 2.** Detailed definitions and examples of intents in the VPED dataset

| **Intent Label** | **Semantic Definition** | **Illustrative Examples (Vietnamese / *English translation*)** |
|---|---|---|
| `add_expense` | Records a newly incurred expense transaction. | "Vừa mua 3 ly trà sữa tốn 75k" (*Just bought 3 bubble teas for 75k*) |
| `update_expense` | Modifies metadata (amount, time, description) of an existing transaction. | "Sửa món phở hôm qua thành cơm tấm" (*Change yesterday's Pho to Broken Rice*) |
| `delete_expense` | Cancels or removes an erroneous or reverted expense record. | "Xoá chi tiêu đi chơi 5tr1 hôm qua" (*Delete the 5.1m trip expense from yesterday*) |
| `search_expense` | Retrieves or searches for past transaction information. | "xem lại tiền ăn sáng nay" (*Review this morning's breakfast cost*) |
| `stat_expense` | Requests data aggregation or statistical calculations by time or category. | "Tổng hợp tuần này tôi đã tiêu bao nhiêu" (*Summarize how much I spent this week*) |

VPED is our developed dataset in the personal expense management domain. It is designed to better capture multi-intent interactions. It contains 1,339 utterances, of which over 50% express multiple intents, spanning 16 slot labels and 5 base intents (refer to Table 2 for a comprehensive list of intent definitions), which expand to 31 compound intents. Data were collected through a questionnaire-based survey targeting Vietnamese speakers who regularly manage personal or family expenses. Participants were asked to simulate natural interactions with a virtual expense assistant, providing example utterances and possible slot values (e.g., expense type, amount, and date). The collected data were classified, annotated, and augmented using oversampling, concatenation, dropout, and entity swapping to increase data diversity and balance intent distributions.

### 4.2 Experiment Setup

We conduct a series of experiments to evaluate the performance of VSLIM on two datasets. Our evaluation focuses on three main aspects: (i) quantitative comparison with multiple baselines across different settings, (ii) the impact of varying intent and slot characteristics, and (iii) the role of token-level intent labels in enhancing overall model performance.

**Table 3.** Results on PhoATIS dataset. $^{mul}$ indicates multi-intent model. Results with $^{r}$ are taken from previous papers. **Bold** numbers are the best results in each column, while the second best is underlined. VSLIM with PhoBERTv1 shows competitive result, even without slot-intent labels.

| Model | PhoATIS | | |
|---|---|---|---|
| | Intent Acc | Slot F1 | SeFr Acc (basic) |
| AGIF$^{mul}$ | 95.63 | 92.42 | 78.42 |
| JointBERT+CRF + PhoBERTv1 | 97.40$^{r}$ | 94.75$^{r}$ | 85.55$^{r}$ |
| JointIDSF + PhoBERTv1 | 97.62$^{r}$ | **94.98**$^{r}$ | **86.25**$^{r}$ |
| MISCA$^{mul}$ + PhoBERTv1 | 97.06 | 94.60 | 84.71 |
| Co-Guiding$^{mul}$ | 95.14 | 93.22 | 80.44 |
| Our VSLIM$^{mul}$ + PhoBERTv1 | **97.80** | 94.75 | 86.09 |

*Model configuration.* VSLIM is built on the PhoBERT-base encoder, which includes 12 transformer layers with 768 hidden units and 12 attention heads. Input sequences are padded or truncated to a maximum length of 128 tokens, and training is carried out with a batch size of 32. We determined the optimal hyperparameters by conducting multiple experiments via randomized search. The search space covered dropout rates in $\{0.1, 0.2, 0.3\}$, batch sizes in $\{16, 32, 64\}$, and loss weights $\alpha, \beta, \gamma$ in $\{0.5, 1.0, 2.0\}$. The configuration reported herein corresponds to the best-performing setup on the validation set based on semantic frame accuracy. We apply a dropout rate of 0.1 to the output of all classifiers. During the training of the Slot-Intent task, we apply teacher forcing but disable it during inference.

*Baselines.* When evaluating single-intent models on the multi-intent task, we adopt the method from [17], where multiple intents are concatenated with a '#' delimiter and treated as a single intent label. All our reported results are the average over 5 runs with highest validation results.

On the PhoATIS dataset, we compare VSLIM against several baselines, including [17], [3], [4], [15] and [25]. Since PhoATIS does not include token-level intent annotations, we use utterance-level intent labels as a proxy, by marking every slots in an utterance with its utterance-level intent label, though this may restrict VSLIM's ability to learn fine-grained slot-intent interactions. JointBERT and JointIDSF have already had results using PhoBERTv1, so we will use PhoBERTv1 as encoder for VSLIM and other models that have competitive results.

On the VPED dataset, we evaluate VSLIM against single and multi-intent models of [17], [4], [15] and [25], using each model's original hyperparameter settings. Here we switch to using latest version of PhoBERTv2 for better results.

**Table 4.** Results on VPED dataset. $^{mul}$ indicates multi-intent model. **Bold** numbers are the best results in each column, with statistically significant improvement ($p < 0.05$ under t-test), while the second best is underlined.

| Model | VPED | | | |
|---|---|---|---|---|
| | Intent Acc | Slot F1 | SeFr Acc basic | SeFr Acc exact |
| AGIF$^{mul}$ | 75.22 | 62.32 | 28.31 | - |
| JointIDSF+PhoBERTv2 | <u>89.74</u> | 85.02 | 54.08 | - |
| MISCA$^{mul}$ + PhoBERTv2 | 89.06 | <u>85.39</u> | <u>54.44</u> | - |
| Co-Guiding$^{mul}$ | 75.59 | 67.05 | 29.66 | - |
| Our VSLIM$^{mul}$+PhoBERTv2 | **91.30** | **87.23** | **58.95** | **56.72** |

### 4.3 Results and Discussion

We evaluate all models using three standard metrics for joint intent detection and slot filling: Intent Accuracy (Intent Acc), Slot Filling F1 (Slot F1), and Semantic Frame Accuracy (SeFr Acc) [24]. For SeFr Acc, we report two variants: (i) *Basic*, which measures correctness at the utterance level based on intent and slot predictions, and (ii) *Exact*, which additionally considers token-level intent prediction. A comparison of the results can be found in Tables 3 and 4.

*Results on PhoATIS.* As presented in Table 3, VSLIM achieves results on par with state-of-the-art single-intent baselines, even without token-level intent annotations. Using PhoBERTv1, it reaches 97.80% Intent Accuracy and 94.75% Slot F1, highlighting its effectiveness under limited supervision.

*Results on VPED.* On the more complex, multi-intent VPED dataset (Table 4), VSLIM outperforms all competing models. It achieves 91.30% Intent Accuracy, 87.23% Slot F1, and 58.95% SeFr Acc (Basic) - a relative improvement of more than +4.5 percentage points in SeFr Acc compared to the second strongest baseline. These gains are largely attributed to the biaffine classifier, which effectively captures slot-intent dependencies, improving both consistency and interpretability in predictions.

Overall, VSLIM performs competitively on single-intent tasks and delivers substantial improvements in multi-intent settings. These results validate the importance of token-level intent modeling and structured slot-intent interaction learning in enhancing joint ID-SF performance.

### 4.4 Ablation Study

To better understand the functionality of key components in VSLIM, we performed an ablation study using two variants of the model:

- **Without biaffine classifier (*w/o biaffine*)**: This variant removes the biaffine classifier in slot-intent mapping task and instead uses a concatenation-based feed-forward layer as in [2].

**Table 5.** Ablation study results on the VPED validation set. **Bold** numbers are best results in each column.

| Model | Intent Acc | Slot F1 | SeFr Acc (Basic) | SeFr Acc (Exact) |
|---|---|---|---|---|
| VSLIM | **97.06** | **95.75** | **83.91** | **83.06** |
| *w/o biaffine* | 95.95 (-1.11) | 94.20 (-1.55) | 81.90 (-2.01) | 80.95 (-2.11) |
| *w/ CRF* | 95.56 (-1.50) | 94.53 (-1.22) | 82.14 (-1.77) | 80.63 (-2.43) |

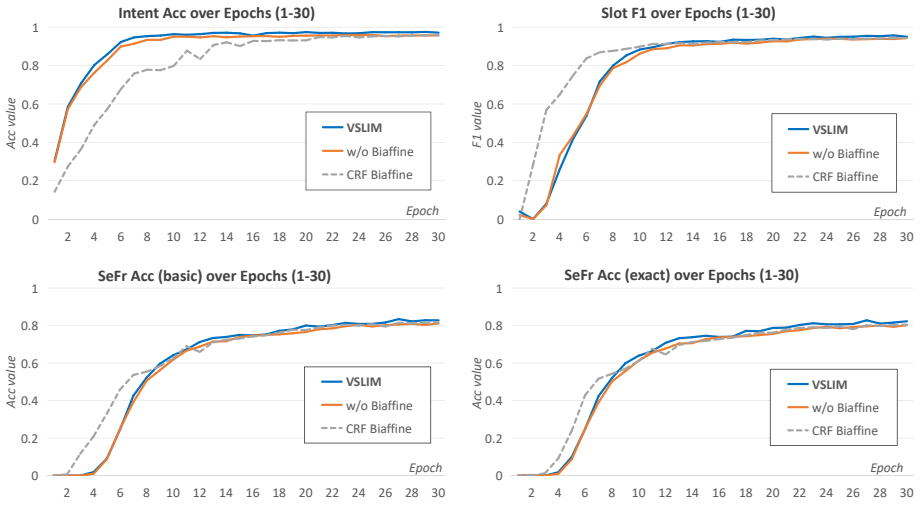


**Fig. 3.** Training dynamics of VSLIM and its ablation variants on the VPED validation set. VSLIM shows faster convergence and higher training stability compared to the variants without the biaffine classifier or with a CRF layer.

– **With CRF (*w/ CRF*)**: This variant adds a Conditional Random Field (CRF) layer for slot prediction to investigate whether structured sequence modeling improves label consistency.

We use the *w/o biaffine* variant to examine the role of biaffine modeling in token-level intent classification. In contrast, the *w/ CRF* variant tests whether capturing sequential dependencies across slot labels improves overall model performance. In the CRF setting, slot representations $y_k^s$ are passed to a linear-chain CRF layer, which models transitions between labels. Validation results on the VPED dataset are shown in Table 5.

Removing the biaffine classifier leads to consistent performance drops across all metrics: Intent Accuracy decreases by **1.11%**, Slot F1 by **1.55%**, and Semantic Frame Accuracy (Basic) by **2.01%**. These results highlight the importance of biaffine modeling for learning slot-intent interactions. Interestingly, introducing a CRF layer also results in slightly lower scores, suggesting that the added sequential constraints may hinder optimization in multi-intent settings.

Figure 3 further illustrates the learning dynamics across variants. VSLIM reaches a validation Intent Accuracy of 0.95 by epoch 8, compared to epochs 10 and 23 for the *w/o biaffine* and *w/ CRF* models, respectively. Although the CRF variant initially improves consistency, all models converge to similar Slot F1 scores (around 0.90) after epoch 12.

In summary, VSLIM demonstrates both higher efficiency and stronger final performance compared to its ablated versions. These findings confirm the value of biaffine classification and suggest that CRF-based sequence modeling may not be always necessary for multi-intent tasks.

## 5 Conclusion

In this paper, we presented VSLIM, a joint, explicit mapping model for multi-intent detection and slot filling in Vietnamese. VSLIM explicitly models the relationship between intents and slots through a biaffine classifier, enhancing the interaction between sentence-level and token-level representations. Experimental results demonstrate that VSLIM achieves competitive performance to existing approaches, while maintaining strong generalization across both single- and multi-intent utterances. Furthermore, although this study focuses on Vietnamese, the VSLIM architecture is inherently language-agnostic. By substituting the encoder with appropriate pre-trained language models and fine-tuning on corresponding annotated datasets, our approach can be readily adapted to other languages, particularly morphologically similar isolating languages or low-resource scenarios where explicit slot-intent mapping is beneficial. This work contributes a strong Vietnamese benchmark and establishes a foundation for further research in multilingual and low-resource joint semantic parsing.

## Acknowledgements

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

## References


1. Bunk, T., Varshneya, D., Vlasov, V., Nichol, A.: DIET: Lightweight language understanding for dialogue systems. arXiv preprint arXiv:2004.09936 (2020)
2. Cai, F., Zhou, W., Mi, F., Faltings, B.: 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). pp. 7607–7611. IEEE (2022)
3. Chen, Q., Zhuo, Z., Wang, W.: BERT for Joint Intent Classification and Slot Filling (2019)
4. Dao, M.H., Truong, T.H., Nguyen, D.Q.: Intent Detection and Slot Filling for Vietnamese. In: Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH) (2021)
5. Dozat, T., Manning, C.D.: Deep Biaffine Attention for Neural Dependency Parsing (2017)
6. Gangadharaiah, R., Narayanaswamy, B.: Joint Multiple Intent Detection and Slot Labeling for Goal-Oriented Dialog. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 564–569. Association for Computational Linguistics (Jun 2019)
7. Goo, C.W., Gao, G., Hsu, Y.K., Huo, C.L., Chen, T.H., Hsu, K.W., Chen, Y.N.: Slot-gated modeling for joint slot filling and intent prediction. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 753–757. Association for Computational Linguistics (2018)
8. Hakkani-Tür, D., Tür, G., Celikyilmaz, A., Chen, Y.N., Gao, J., Deng, L.: Multi-Domain Joint Semantic Frame Parsing using Bi-directional RNN-LSTM. In: Proceedings of Interspeech (2016)
9. Kim, Y.: Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics (Oct 2014)
10. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270. Association for Computational Linguistics (2016)
11. Liu, B., Lane, I.: Attention-based recurrent neural network models for joint intent detection and slot filling. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH). pp. 685–689. ISCA (2016)
12. Mesnil, G., Dauphin, Y.N., Yao, K., Bengio, Y.: Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**, 530–539 (2015)
13. Nguyen, D.Q., Tuan Nguyen, A.: PhoBERT: Pre-trained language models for Vietnamese. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1037–1042. Association for Computational Linguistics (2020)
14. Nguyen, D.Q., Verspoor, K.: End-to-end neural relation extraction using deep biaffine attention. In: European conference on information retrieval. pp. 729–738. Springer (2019)

15. Pham, T., Tran, C., Nguyen, D.Q.: MISCA: A Joint Model for Multiple Intent Detection and Slot Filling with Intent-Slot Co-Attention. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 12641–12650. Association for Computational Linguistics (Dec 2023)
16. Price, P.: Evaluation of spoken language systems: The ATIS domain. In: Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990 (1990)
17. Qin, L., Xu, X., Che, W., Liu, T.: AGIF: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1807–1816. Association for Computational Linguistics (Nov 2020)
18. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Natural language processing using very large corpora, pp. 157–176. Springer (1999)
19. Trang, N.T.T., Anh, D.T.D., Viet, V.Q., Woomyoung, P.: Advanced joint model for Vietnamese intent detection and slot tagging. In: International Conference on Industrial Networks and Intelligent Systems. pp. 125–135. Springer (2022)
20. Tu, N.A., Hieu, D.X., Phuong, T.M., Bach, N.X.: A bidirectional joint model for spoken language understanding. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
21. Vu, N.T.: Sequential convolutional neural networks for slot filling in spoken language understanding. arXiv preprint arXiv:1606.07783 (2016)
22. Wang, D., Ni, Q.: A domain-aware model with multi-perspective contrastive learning for natural language understanding. Applied Intelligence (2025)
23. Wang, Y., Shen, Y., Jin, H.: A Bi-Model Based RNN Semantic Frame Parsing Model for Intent Detection and Slot Filling. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 309–314. Association for Computational Linguistics (Jun 2018)
24. Weld, H., Huang, X., Long, S., Poon, J., Han, S.C.: A survey of joint intent detection and slot filling models in natural language understanding. ACM Computing Surveys **55**, 1–38 (2022)
25. Xing, B., Tsang, I.W.: Co-guiding for multi-intent spoken language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**, 2965–2980 (2023)
26. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for Document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. pp. 1480–1489 (2016)
27. Yu, J., Bohnet, B., Poesio, M.: Named Entity Recognition as Dependency Parsing. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6470–6476. Association for Computational Linguistics (Jul 2020)
28. Zhang, C., Li, Y., Du, N., Fan, W., Yu, P.: Joint Slot Filling and Intent Detection via Capsule Neural Networks. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 5259–5267 (2019)
29. Zhu, Z., Xu, W., Cheng, X., Song, T., Zou, Y.: A Dynamic Graph Interactive Framework with Label-semantic Injection for Spoken Language Understanding. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)