Topology-Enhanced Experts for Continuous Intent Trigger Modeling in Disentangled SLU

1st Chuanwang Xiong
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, Jiangsu, China

2nd Hongwei Ge*
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, Jiangsu, China

***Abstract*—Multi-intent Spoken Language Understanding (SLU) is critical for deciphering complex user utterances in real-world dialogue systems. However, current methods are constrained by two fundamental limitations: the entanglement of multi-granularity semantics leading to feature collapse, and the indiscriminate treatment of tokens which ignores the structural locality of intent triggers. To address these challenges, we propose TECT (Topology-enhanced Experts for Continuous Triggers), a unified framework that injects explicit inductive biases for granular disentanglement and structural awareness. Our approach introduces two key mechanisms: 1) Orthogonal Multi-view Experts (OME), which integrates utterance-, chunk-, and token-level semantics. Crucially, OME imposes a rigorous pairwise orthogonality constraint to compel experts to capture complementary, non-redundant representations within the semantic manifold; and 2) Topology-Aware Trigger Selection (TATS), which redefines the selection process to identify coherent semantic spans. TATS incorporates a relative positional advantage bias to leverage topological structure and employs adaptive sparsity to dynamically modulate the focus area based on intent complexity. Extensive experiments on the MixATIS and MixSNIPS benchmarks demonstrate that TECT establishes new state-of-the-art performance, highlighting the superior efficiency and robustness of our structural design in capturing multi-granularity linguistic patterns.**

***Index Terms*—Intent Detection, Slot Filling, Multi-view learning, Spoken Language Understanding**

I. Introduction

Spoken Language Understanding (SLU) serves as the cornerstone of task-oriented dialogue systems, transforming user utterances into actionable semantic frames. While early research focused on simple single-intent scenarios, real-world interactions often involve complex, multi-intent expressions (e.g., *"Book a flight to London and find a hotel nearby"*). Consequently, Multi-intent SLU has emerged as a pivotal research frontier, aiming to simultaneously identify multiple intents and extract corresponding slots. As conversational agents evolve toward greater autonomy and complexity, the ability to robustly disentangle and interpret these intricate semantic structures is becoming increasingly critical for next-generation dialogue systems.

Despite recent advancements, existing approaches remain constrained by two fundamental limitations that hinder their effectiveness in complex scenarios. **First, the ineffective fusion and entanglement of multi-granularity semantics.** Robust SLU requires the integration of diverse semantic views—utterance-level global context, chunk-level local phrases, and token-level fine-grained details—to capture a complete picture of user intent. However, prior methods often fail to exploit the complementarity of these views. Monolithic models typically rely on a single dominant view, while multi-scale approaches like CLID [1] and Uni-SLU [2] simply concatenate features without explicit constraints. This leads to feature collapse, where experts at different levels degenerate into learning homogeneous, redundant representations. Without mechanisms to enforce distinctiveness, the model cannot effectively leverage the unique strengths of each granularity, resulting in a suboptimal and entangled feature space.

Second, the indiscriminate treatment of tokens and loss of structural locality. In multi-intent utterances, intent triggers typically manifest as **continuous semantic spans** (e.g., the phrase *"find a hotel"*) rather than isolated keywords. However, most current methods, such as Stack-Propagation [3] employ standard attention mechanisms that treat the sentence as a "bag of words." This approach suffers from two critical flaws: (1) It ignores the topological locality of triggers, failing to capture the continuous structural nature of intent regions; (2) It lacks an adaptive scope, unable to determine the appropriate attention span in multi-intent scenarios. Consequently, critical intent signals are often diluted by background noise or fragmented into disjointed points.

To address these limitations, we propose TECT (**T**opology-enhanced **E**xperts for **C**ontinuous **T**riggers), a unified framework that injects explicit inductive biases to achieve granular disentanglement and structural awareness. First, we introduce the Orthogonal Multi-view Experts (OME) module to resolve feature entanglement. OME incorporates experts operating at three distinct semantic levels: *Utterance View*, *Chunk View*, and *Token View*. Instead of naive fusion, we strictly enforce a Pairwise Orthogonality Constraint across all three views. By minimizing the mutual information between every pair of experts, we mathematically compel each view to capture a unique, non-redundant subspace of the semantic manifold. This ensures true complementarity, preventing feature collapse and maximizing the information gain from each granularity. Building on this, we propose the Topology-Aware Trigger Selection (TATS) mechanism to address the "bag-of-words"

*Corresponding author.

limitation. TATS redefines attention by incorporating a relative positional advantage bias, utilizing Gaussian convolution to respect the continuous topological structure of intent triggers. Furthermore, it features adaptive sparsity, where the attention scope is dynamically modulated by an auxiliary intent counter. This allows the model to "zoom in" on the precise number of key semantic regions required—effectively filtering noise while preserving the integrity of intent spans. Extensive experiments on the MixATIS and MixSNIPS benchmarks demonstrate that TECT establishes new state-of-the-art performance. Our results highlight that by injecting explicit inductive biases for multi-granularity disentanglement and structural locality, TECT achieves superior robustness and precision in complex multi-intent SLU tasks.

The main contributions of this work are as follows:

- A novel unified framework, termed TECT, is proposed to address feature entanglement and structural neglect in multi-intent SLU by injecting explicit inductive biases.
- The Orthogonal Multi-view Experts (OME) module is introduced, which employs a rigorous pairwise orthogonality constraint to compel utterance-, chunk-, and token-level experts to capture complementary, disentangled semantics.
- The Topology-Aware Trigger Selection (TATS) mechanism is designed to robustly identify continuous intent semantic spans, leveraging Gaussian-smoothed locality propagation and adaptive sparsity.

II. Methodology

We propose TECT, a framework designed to disentangle semantic features for multi-intent detection and slot filling (Fig. 1). The workflow begins with a Shared Encoder, followed by the Orthogonal Multi-view Experts (OME) module which extracts distinct granular features. These features are dynamically fused before passing to task-specific decoders.

A. Problem Definition

Let $x = (x_1, \ldots, x_n)$ be an input utterance. The goal is twofold: (1) predict a set of intent labels o^I and (2) assign a semantic label o_i^S to each token x_i .

B. Shared Encoder

We utilize BERT [4] as the backbone encoder. Given x, the encoder yields a sequence of contextualized hidden states $\mathbf{h} = (h_1, \ldots, h_n)$, where $h_i \in \mathbb{R}^{d_{model}}$. Unlike standard approaches relying solely on the [CLS] token, we leverage the full sequence $\mathbf{h}$ to preserve fine-grained semantic details.

C. Orthogonal Multi-view Experts (OME)

To resolve feature entanglement, we introduce the Orthogonal Multi-view Experts (OME) module. OME incorporates experts operating at three distinct semantic levels: *Utterance View*, *Chunk View*, and *Token View*. Instead of naive fusion, we strictly enforce a **Pairwise Orthogonality Constraint** across these views. By minimizing the mutual information between experts, we mathematically compel each view to capture a unique, non-redundant subspace of the semantic manifold, preventing feature collapse and maximizing information gain.

a) Multi-granularity Feature Extraction: We employ a set of parallel experts for each view to extract diverse features:

- **Utterance View (H^{ut}):** Focuses on global semantic coherence. We project $\mathbf{h}$ into a compact space, apply pooling to obtain a global vector $\mathbf{h}^{ut}$, and process it through an MLP with residual connections. The output is broadcasted to match the sequence length.
- **Chunk View (H^{ck}):** Captures local dependencies. We utilize a sliding window of size k to aggregate neighboring token information. For each token position t, a local context window $H_t \in \mathbb{R}^{k \times d_{model}}$ is weighted and summed to form the chunk representation, effectively modeling phrase-level structures.
- **Token View (H^{tk}):** Refines fine-grained details. Each token embedding is independently processed via non-linear transformations to enhance discriminative power at the word level.

b) Task-Adaptive Gating Fusion: Acknowledging that intent detection prioritizes global signals while slot filling relies on local details, we employ a task-adaptive gating strategy. For a specific task (e.g., Intent I), we compute a gate distribution W^I over all experts based on the global sentence representation. The final multi-view representation I^I is a weighted integration of the expert outputs:

$$W^I = \text{Softmax}(\text{MLP}^I(\text{Pooling}(\mathbf{h}))) \tag{1}$$

$$I^I = \sum_{v \in \{ut,ck,tk\}} \sum_{j=1}^{n_v} W_j^I \cdot H_j^v \tag{2}$$

where n_v denotes the number of experts for view v. A separate gate W^S is computed analogously to generate I^S for the slot filling task.

To prevent feature collapse and ensure the distinctiveness of each view, we impose an orthogonality constraint $\mathcal{L}_{orth}$. This regularizer minimizes the cosine similarity between the representations of different views (e.g., maximizing the angle between global H^{ut} and H^{tk}vectors), thereby forcing experts to capture complementary semantic information:

$$\mathcal{L}_{orth} = \sum_{v1,v2 \in \{ut,ck,tk\}} \left\| \frac{H^{v1} \cdot (H^{v2})^T}{\|H^{v1}\| \|H^{v2}\|} \right\|_F^2 \tag{3}$$

D. Auxiliary Task: Intent Counting

To assist in multi-label decoding, we incorporate an auxiliary regression task to predict the number of intents o^a. Crucially, this prediction o^a not only serves as a final ranking threshold but also acts as an adaptive sparsity controller for the TATS module. We utilize the aggregated intent features I^I:

$$\hat{y}^a = \text{Softmax}(\text{MLP}(\text{Pooling}(I^I))) \tag{4}$$

$$o^a = \text{argmax}(\hat{y}^a) \tag{5}$$

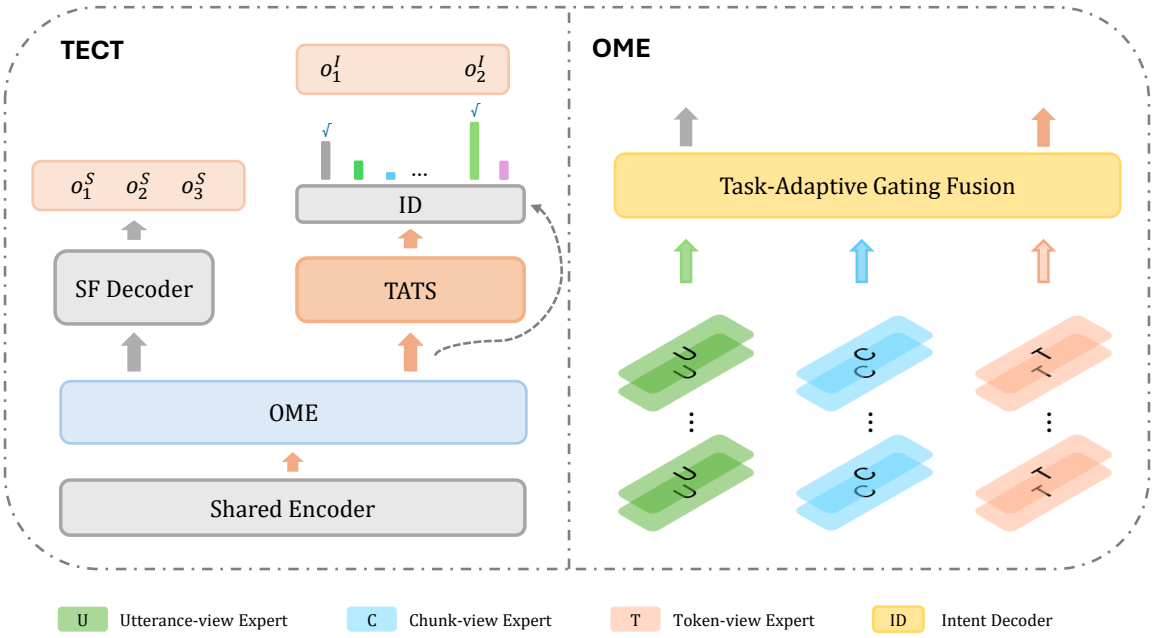

Fig. 1. Overall architecture of TECT

E. Topology-Aware Trigger Selection (TATS)

Prior studies typically treat all tokens uniformly, introducing noise from irrelevant words like stopwords. Standard attention mechanisms exacerbate this by modeling the sentence as a "bag of words," ignoring the topological locality of intent triggers. Furthermore, they lack the adaptability to determine the appropriate attention span for multi-intent scenarios. To address these limitations, we propose an enhanced Topology-Aware Trigger Selection mechanism, which incorporates relative positional advantage and adaptive sparsity.

First, we compute the raw semantic relevance score $S_{sem} \in \mathbb{R}^n$ for each token using a learnable query vector w_q:

$$S_{sem} = I^I \cdot w_q^T \tag{6}$$

To capture the topological structure of intent triggers—premised on the inductive bias that intent-bearing tokens typically exhibit spatial clustering—we introduce a *Gaussian-Smoothed Locality Propagation* mechanism. Rather than treating tokens as isolated entities, we model the intent distribution as a continuous field. We employ a learnable Gaussian kernel to diffuse the semantic salience S_{sem} across local neighborhoods. This process functions as a spatial density estimator, allowing highly indicative tokens to support their neighbors and forming coherent intent clusters:

$$S_{loc} = \mathcal{G}(S_{sem}) \tag{7}$$

where $\mathcal{G}(\cdot)$ represents the parameterized Gaussian operation. The final trigger importance S_{final} is derived by fusing the intrinsic semantic confidence with this structural locality bias:

$$S_{final} = (1 - \lambda) \cdot S_{sem} + \lambda \cdot S_{loc} \tag{8}$$

where λ balances the contribution of semantic precision and topological coherence.

Next, to handle multi-intent complexity, we employ adaptive sparsity controlled by the auxiliary task output o^a. We hypothesize that the number of tokens to select should be proportional to the number of intents. Thus, we generate a dynamic mask M_{dyn} that only retains the top-K tokens, where K is a function of the predicted intent count o^a:

$$M_{dyn}[i] = \begin{cases} 0 & \text{if } S_{final}[i] \in \text{TopK}(S_{final}) \\ -\infty & \text{otherwise} \end{cases} \tag{9}$$

The final attention weights α are obtained by applying Softmax on the masked scores:

$$\alpha = \text{Softmax}(S_{final} + M_{dyn} + \mathcal{M}_{pad}) \tag{10}$$

where $\mathcal{M}_{pad}$ masks padding tokens. The focused intent representation v_{intent} is then computed:

$$v_{intent} = \sum_{i=1}^{n} \alpha_i \cdot I_i^I \tag{11}$$

Finally, the intent probabilities $\hat{y}^I$ are predicted from this focused vector:

$$\hat{y}^I = \text{Sigmoid}(\text{MLP}(v_{intent})) \tag{12}$$

During inference, we select the top-o^a intents from $\hat{y}^I$ as the final prediction o^I.

F. Aligned Slot-Filling Decoder

To ensure strict alignment between input tokens and generated slot labels, we adopt the aligned decoder architecture proposed by [5]. This method incorporates a zero-diagonal

mask M into the cross-attention mechanism, restricting the decoder to attend exclusively to the corresponding t-th position of the encoder output during the t-th generation step. Taking the multi-view fusion information I^S as input, the final slot probability distribution $\hat{y}^S$ and the predicted sequence o^S are derived as follows:

$$\hat{y}^S = \text{Softmax}\left(\text{MLP}\left(\text{Dec}(I^S, M)\right)\right) \tag{13}$$

$$o^S = \text{argmax}\left(\hat{y}^S\right) \tag{14}$$

where $\text{Dec}(\cdot)$ denotes the stacked decoder layers with the diagonal masking constraint.

G. Joint Training with Orthogonality Constraint

We define the training objective $\mathcal{L}_I$ of intent detection:

$$\mathcal{L}_I = \sum_{j=1}^{n} \sum_{i=1}^{n_I} CE(y_{i,j}^I,\ \mathbf{o}_{i,j}^I) \tag{15}$$

$$CE(y, \hat{y}) = y\ log(\hat{y}) + (1-y)\ log(1-\hat{y}) \tag{16}$$

where $y_{i,j}^I$ is the gold intent label, n_I denotes the number of intent labels.

Similarly, the task objective of slot filling $\mathcal{L}_S$ is formulated as:

$$\mathcal{L}_S = -\sum_{j=1}^{n} \sum_{i=1}^{n_S} y_{i,j}^S\ log(\mathbf{o}_{i,j}^S) \tag{17}$$

where $y_{i,j}^S$ is the gold slot label, n_S denotes the number of slot labels.

The final training objective $\mathcal{L}$ is a weighted sum of task losses and the regularization term:

$$\mathcal{L} = \alpha\,\mathcal{L}_I + (1-\alpha)\mathcal{L}_S + \lambda_1\,\mathcal{L}_c + \lambda_2\,\mathcal{L}_{orth} \tag{18}$$

where λ_1, λ_2 and α are hyper-parameters to balance the tasks, and $\mathcal{L}_c$ denotes the cross-entropy loss for the auxiliary task.

III. Experiments

A. Data

To comprehensively evaluate the proposed model, we conducted experiments on two representative multi-intent Spoken Language Understanding (SLU) datasets: MixATIS [9], [16] and MixSNIPS [9], [16]. MixATIS is a multi-intent version derived from the classic ATIS [17] dataset, with its domain focused on airline information queries. It comprises 13,162 training samples, 756 validation samples, and 828 testing samples. MixSNIPS [18], an extension of the SNIPS dataset, is a multi-intent, multi-domain dataset covering areas such as hotels, restaurants, and movies. It consists of 39,776 training samples, 2,198 validation samples, and 2,199 testing samples. Notably, sentences in both multi-intent SLU datasets often contain multiple intents, with the distribution of samples containing 1, 2, and 3 intents being 30%, 50% and 20% respectively.

B. Experimental Settings

Optimization is performed using the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We standardize the batch size at 80 for both datasets, while setting the maximum sequence lengths to 80 for MixATIS and 60 for MixSNIPS. Regarding the expert configuration in OME, the MixATIS dataset utilizes 3 utterance experts, 1 token expert, and 5 chunk experts. Conversely, for MixSNIPS, both utterance and token experts are set to 2, with chunk experts remaining at 5. Hyperparameters were empirically optimized based on performance on the development set. This specific configuration reflects the linguistic complexity variance between the two datasets. To mitigate overfitting, we implement an early-stopping mechanism with a patience of 10 epochs (monitoring the overall loss $\mathcal{L}$ on the development set). The reported results represent the average of several independent runs. All experiments were conducted on an NVIDIA 4090D GPU.

C. Main Results and Metrics

To comprehensively assess model performance, we employ three standard metrics: F1 score for Slot Filling (Slot F1), Accuracy for Intent Detection (Intent Acc), and Overall Accuracy (Overall Acc). Among these, Overall Acc serves as the strictest criterion, as it considers a prediction correct only if both the intent and all slots within an utterance are perfectly recognized.

Table I presents the comparative results on the MixATIS and MixSNIPS benchmarks. On the MixATIS dataset, our model achieves state-of-the-art performance with an Intent Acc of 84.5 and an Overall Acc of 53.7. The latter metric highlights the robustness of our approach compared to existing baselines. For instance, TECT outperforms CLID [1] and Uni-MIS [2] by significant margins, exceeding their Overall Acc scores by relative margins of 4.7% and 1.2%, respectively. Similarly, on the MixSNIPS dataset, the proposed model demonstrates distinct superiority across all metrics. It achieves a Slot F1 of 97.1, edging out Uni-MIS (96.4), and an Intent Acc of 97.9, surpassing MISCA (97.3). Most notably, in terms of Overall Acc, our model reaches 86.7, marking a substantial 3.9% relative improvement over Uni-MIS (83.4). These consistent gains validate the efficacy of the proposed Orthogonal Multi-view Experts strategy, confirming its ability to integrate multi-granularity information effectively for complex SLU tasks.

IV. Analysis

To investigate the individual contributions of the core components in TECT, we conducted ablation experiments on the MixATIS and MixSNIPS datasets. The results are summarized in Table II.

A. Impact of Orthogonal Multi-view Experts (OME)

By removing the OME module, the model relies on standard shared embeddings without multi-granularity disentanglement. On MixATIS, we observe a significant drop in Intent Accuracy (7.0%) and Overall (ALL) Accuracy (8.1%). On MixSNIPS,

Model Type	Model	MixATIS			MixSNIPS		
		Slot F1(%)	Intent Acc(%)	ALL Acc(%)	Slot F1(%)	Intent Acc(%)	ALL Acc(%)
Conventional Model	Bi-Model [6]	83.9	70.3	34.4	90.7	95.6	63.4
	Slot-Gated [7]	87.7	63.9	35.5	87.9	94.6	55.4
	SF-ID Network [8]	87.4	63.9	35.5	90.6	95.0	59.9
	Stack-Propagation [3]	87.8	72.1	40.1	94.2	96.0	72.9
	AGIF [9]	86.9	72.2	39.2	93.8	95.1	72.7
	GL-GIN [9]	87.2	75.6	41.6	93.7	95.2	72.4
	SDJN [10]	88.2	77.5	49.0	94.3	96.6	75.0
	CLID [1]	88.2	77.5	49.0	94.3	96.6	75.0
	CLID(Roberta)	85.9	80.5	49.4	96.0	97.0	82.2
	Co-guiding Net [11]	89.9	79.1	51.3	95.1	97.7	77.5
	Aligner^2 [12]	**90.6**	78.7	52.8	95.3	98.1	76.5
	MISCA [13]	90.5	76.7	53.0	95.2	97.3	77.9
	Uni-MIS [2]	88.3	78.5	52.5	96.4	97.2	83.4
LLM	EN-Mistral [14]	88.7	80.6	53.4	95.6	97.6	79.8
	DSCP [15]	-	50.7	-	-	89.7	-
	Inter-DSCP [15]	-	52.2	-	-	92.0	-
Conventional Model	TECT(Ours)	88.9	**84.5**	**53.7**	**97.1**	**97.9**	**86.7**

TABLE I
SLU PERFORMANCE ON MIXATIS AND MIXSNIPS DATASETS, THE OPTIMAL EXPERIMENTAL RESULTS ARE SHOWN IN BOLD. CLID(ROBERTA) DENOTES A ROBERTA BACKBONE MODEL OF CLID.

Model	Slot F1(%)	Intent Acc(%)	ALL Acc(%)
MixATIS			
w/o OME	86	77.5	45.6
w/o TATS	87.8	80.4	50.6
w/o TATS & OME	70.4	72.1	32.5
TECT(Ours)	**88.9**	**84.5**	**53.7**
MixSNIPS			
w/o OME	94.4	93.1	73.4
w/o TATS	96.5	95	83.3
w/o TATS & OME	80.6	82	43.2
TECT(Ours)	97.1	**97.9**	**86.7**

TABLE II
ABLATION EXPERIMENTS ON THE MIXATIS AND MIXSNIPS DATASETS. THE OPTIMAL RESULTS ARE SHOWN IN BOLD.

the performance also degrades, particularly for ALL Accuracy which falls to 73.4%. This decline validates that the pairwise orthogonality constraint is essential for resolving feature entanglement, as it compels the model to capture non-redundant semantic information across utterance, chunk, and token granularities. Without this module, the model fails to maintain the distinctiveness of multi-view features, leading to representation collapse in complex multi-intent scenarios.

B. Effectiveness of Topology-Aware Trigger Selection

Excluding the Topology-Aware Trigger Selection mechanism (w/o TATS) results in a consistent decrease in Intent Accuracy and ALL Accuracy (e.g., a 4.1% and 3.4% drop in ALL Acc on MixATIS and MixSNIPS, respectively). This indicates that the Gaussian-Smoothed Locality Propagation is vital for intent detection. By omitting the structural locality bias, the model treats intent-bearing tokens as isolated points rather than cohesive topological regions. This loss of spatial inductive bias makes it harder for the decoder to aggregate intent signals from clustered keywords, thereby harming the overall joint task performance.

C. Joint Contribution of OME and TATS

The most drastic performance collapse occurs when both the OME and TATS modules are removed. On MixATIS, Slot F1 drops by 18.5% and Intent Accuracy by 12.4%, while on MixSNIPS, the ALL Accuracy plummeted by more than half to 43.2%. This sharp decline highlights the synergy between our proposed components. While OME provides the high-quality, disentangled features, TATS ensures those features are strategically utilized based on their topological importance. The cumulative drop proves that the robustness of TECT stems from its ability to both extract non-redundant semantic information and model the spatial dependencies of linguistic triggers simultaneously.

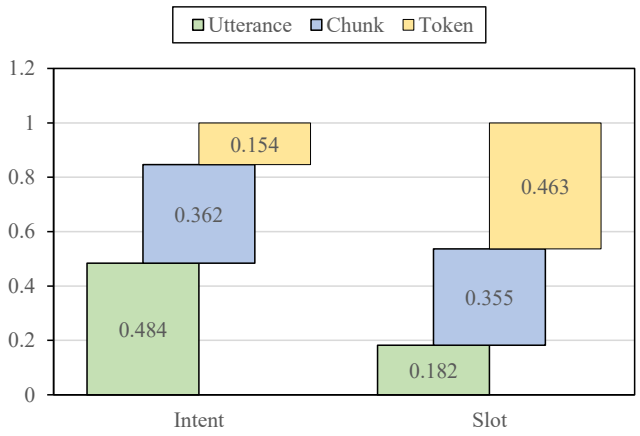

Fig. 2. Distribution of average gating weights for Intent Detection and Slot Filling tasks across three semantic views on the MixATIS dataset. Values indicate the percentage contribution of each semantic granularity to the final task representation

D. Task-Specific View Preference Analysis

To further investigate how the Task-Adaptive Gating mechanism distributes information across the three semantic granularities, we visualize the average gating weights for both tasks on the test set. As shown in Fig. 2, the experimental results exhibit a distinct specialization in feature utilization:

Intent Detection predominantly draws from the Utterance View (48.4%) and Chunk View (36.2%), with only a minimal reliance on the Token View (15.4%). This distribution aligns with the linguistic nature of intent, which typically emerges from global sentence-level semantics or phrase-level local structures rather than individual word properties.

Slot Filling, conversely, shifts its focus toward the Token View (46.3%) and Chunk View (35.5%). Since slot filling requires precise boundary identification and fine-grained labeling for each token, the model naturally prioritizes the word-level and local-context representations provided by these experts.

The sharp contrast in weight distribution (e.g., the Utterance View contribution drops from 48.4% in Intent to 18.2% in Slot) empirically proves that our Orthogonal Multi-view Experts (OME) have indeed captured non-redundant, level-specific features. Moreover, it demonstrates that the task-adaptive gate successfully routes the most relevant semantic information to the corresponding decoder, effectively mitigating task interference in the joint learning framework.

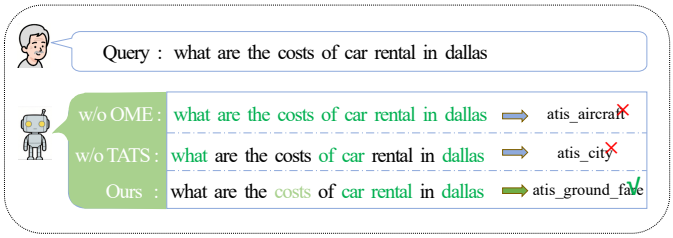

Fig. 3. Case study visualization. Tokens identified as high-priority are highlighted in green, while less relevant tokens are shown in gray.

E. Case Study

To provide a qualitative perspective, we visualize the model's attention distribution in Fig. 3. Green highlights denote tokens identified as salient, while gray indicates background context.

The task requires identifying multiple user intents from a single utterance. Conventional approaches (labeled "w/o TATS") process tokens uniformly, introducing noise that leads to erroneous predictions. When OME is removed but TATS is retained ("w/o OME"), the model attempts to select important tokens but fails due to a lack of contextual grounding. For instance, the model over-weights the token 'dallas', biasing the prediction incorrectly towards 'atis_city'. In contrast, the full TECT framework demonstrates the synergistic integration of its components: OME constructs a comprehensive semantic landscape, enabling TATS to precisely lock onto the critical evidence required for accurate intent classification.

V. Related Work

A. Conventional and Neural SLU Architectures

The intrinsic interdependence between intent detection and slot filling has long motivated the development of joint modeling frameworks. Early research [19] primarily focused on single-intent scenarios. For instance, Hakkani-Tür et al. [20] proposed a bidirectional RNN-LSTM architecture with shared layers to capture dependencies between the two tasks. Similarly, Liu and Lane [21] introduced attention mechanisms to align slot labels with intent-critical tokens.

However, real-world user utterances often contain complex, overlapping goals, shifting the research focus toward multi-intent understanding. Kim et al. [22] addressed this by categorizing sentences and employing a two-stage processing approach. Subsequent works, such as Gangadharaiah and Narayanaswamy [23], utilized intent context vectors for joint prediction. More recently, graph-based frameworks like AGIF [16], GL-GIN [9] and [24] have been developed to explicitly model intent-slot interactions. Additionally, CLID [1] adopted a sliding window-based self-attention scheme to detect span-level intents, while Chen et al. [10] utilized a self-distillation loop for mutual task guidance. Despite these advancements, most existing methods such as [25] and [26], rely heavily on single-view representations or rigid coupling mechanisms. This limitation hinders their ability to effectively decouple task-specific features from noisy contexts. Our work addresses this gap by introducing a multi-view fusion framework equipped with adaptive token weighting to enhance robustness in multi-intent scenarios.

B. Large Language Models in SLU

The emergence of Large Language Models (LLMs) has significantly impacted the field of SLU. With their extensive pre-training, LLMs demonstrate exceptional capabilities in context comprehension and instruction following. Recent studies [14] indicate that with carefully designed prompts, LLMs can directly generate structured intent and slot outputs in a single inference pass, simplifying the development pipeline. In multi-intent settings, LLMs excel at grasping global semantics. For example, DSCP [15] decomposes multi-intent detection into explicit stages—division, solution, and combination—to improve interpretability and reasoning tracking.

While LLMs offer a powerful paradigm for handling open-ended queries, they often function as monolithic "black boxes" that lack fine-grained control over specific structural extractions without extensive resource consumption. In contrast, our approach aims to achieve comparable or superior performance, particularly in low-resource settings, by leveraging a specialized, data-efficient architecture that explicitly models multi-granularity features.

VI. Conclusion

In this paper, we introduced TECT, a unified framework designed to address the challenges of feature entanglement and structural neglect in multi-intent SLU. By synergizing

Orthogonal Multi-view Experts (OME) and **Topology-Aware Trigger Selection (TATS)**, TECT explicitly decouples multi-granularity task representations. The OME module effectively prevents feature collapse by aggregating utterance-, chunk-, and token-level semantics under a rigorous orthogonality constraint. Simultaneously, TATS leverages relative positional advantages and adaptive sparsity to dynamically amplify intent-relevant spans based on their structural locality. Extensive experiments on the MixATIS and MixSNIPS benchmarks demonstrate that TECT significantly outperforms existing state-of-the-art baselines. These results validate that injecting explicit inductive biases for granular disentanglement and topological awareness is highly effective for interpreting complex semantic structures. Future research will explore the cross-lingual and cross-domain transferability of the proposed orthogonal multi-view mechanisms.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grants 52374155, Anhui Provincial Natural Science Foundation under Grant No. 2308085MF218, and PAPD of Jiangsu Higher Education Institutions.

REFERENCES

[1] H. Huang, P. Huang, Z. Zhu, J. Li, and P. Lin, "Clid: A chunk-level intent detection framework for multiple intent spoken language understanding," *IEEE Signal Processing Letters*, vol. 29, pp. 2123–2127, 2022.

[2] S. Yin, P. Huang, and Y. Xu, "Uni-mis: United multiple intent spoken language understanding via multi-view intent-slot interaction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 19 395–19 403.

[3] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "A stack-propagation framework with token-level intent detection for spoken language understanding," 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)*, Nov. 2019, pp. 2078–2087.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 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)*, 2019, pp. 4171–4186.

[5] M. Rafiepour and J. S. Sartakhti, "Ctran: Cnn-transformer-based network for natural language understanding," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107013, 2023.

[6] Y. Wang, Y. Shen, and H. Jin, "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)*, Jun. 2018, pp. 309–314.

[7] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, "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, Volume 2 (Short Papers)*, 2018, pp. 753–757.

[8] H. E, P. Niu, Z. Chen, and M. Song, "A novel bi-directional interrelated model for joint intent detection and slot filling," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Jul. 2019, pp. 5467–5471.

[9] L. Qin, F. Wei, T. Xie, X. Xu, W. Che, and T. Liu, "GL-GIN: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Aug. 2021, pp. 178–188.

[10] L. Chen, P. Zhou, and Y. Zou, "Joint multiple intent detection and slot filling via self-distillation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7612–7616.

[11] B. Xing and I. Tsang, "Co-guiding net: Achieving mutual guidances between multiple intent detection and slot filling via heterogeneous semantics-label graphs," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 159–169.

[12] Z. Zhu, X. Cheng, Y. Li, H. Li, and Y. Zou, "Aligner2: enhancing joint multiple intent detection and slot filling via adjustive and forced cross-task alignment." AAAI Press, 2024.

[13] T. Pham, C. Tran, and D. Q. Nguyen, "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*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 641–12 650.

[14] S. Yin, P. Huang, Y. Xu, H. Huang, and J. Chen, "Do large language model understand multi-intent spoken language?" *arXiv preprint arXiv:2403.04481*, 2024.

[15] L. Qin, Q. Chen, J. Zhou, J. Wang, H. Fei, W. Che, and M. Li, "Divide-solve-combine: An interpretable and accurate prompting framework for zero-shot multi-intent detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 25 038–25 046.

[16] L. Qin, X. Xu, W. Che, and T. Liu, "AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Nov. 2020, pp. 1807–1816.

[17] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The atis spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June, 1990*, 1990, pp. 24–27.

[18] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[19] Y. I. Muhammad, N. Salim, and A. Zainal, "Joint intent detection and slot filling with syntactic and semantic features using multichannel cnn-bilstm," *PeerJ Computer Science*, vol. 10, p. e2346, 2024.

[20] D. Hakkani-Tür, G. Tür, A. Celikyilmaz, Y.-N. Chen, J. Gao, L. Deng, and Y.-Y. Wang, "Multi-domain joint semantic frame parsing using bi-directional rnn-lstm." in *Interspeech*, 2016, pp. 715–719.

[21] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Interspeech 2016*, 2016, pp. 685–689.

[22] B. Kim, S. Ryu, and G. G. Lee, "Two-stage multi-intent detection for spoken language understanding," *Multimedia Tools and Applications*, vol. 76, pp. 11 377–11 390, 2017.

[23] R. Gangadharaiah and B. Narayanaswamy, "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)*, 2019, pp. 564–569.

[24] X. Li, L. Zhang, L. Fang, and P. Cao, "Multi-intent natural language understanding framework for automotive applications: A heterogeneous parallel approach," *Applied Sciences*, vol. 13, no. 17, p. 9919, 2023.

[25] X. Zhuang, X. Cheng, and Y. Zou, "Towards explainable joint models via information theory for multiple intent detection and slot filling," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 17, 2024, pp. 19 786–19 794.

[26] Y. Wang, J. Xu, M. Wang, T. Li, and X. Qu, "A unified prompt-based framework with label semantic expansion for joint multi-intent detection and slot filling." *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 9, pp. 1–14, 2025.