StrucNID: LLM-Guided Structural Prior for Disentangled New Intent Discovery

Jiebin Huang[1], Jindian Su[1*], Ruiqi Wang[1], Xiaobin Ye[2], and Dandan Ma[2]
[1] School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
[2] Guangdong Unicom, Guangzhou, China

***Abstract*—New Intent Discovery (NID) seeks to recognize both known and novel user intents from unlabeled utterances, serving as a fundamental capability for dialogue systems operating in dynamic environments. Existing methods confront two critical challenges, notably Semantic Entanglement, characterized by monolithic embeddings conflating intent with stylistic variations to cause intra-class scatter, and Structural Deficiency, stemming from the lack of topological guidance in LLM-Enhanced approaches that leaves boundary samples ambiguous. To address these issues, we propose StrucNID, a framework explicitly designed for structural disentanglement. We leverage LLMs to construct fine-grained Semantic Mini-Clusters as a binary adjacency graph, providing reliable topological constraints rather than simple data augmentation. We then enforce strict orthogonality between intent and instance subspaces via dual-stream projection, mathematically separating class-invariant semantics from stylistic noise. Finally, we design a Spatio-Temporal voting mechanism that fuses temporal momentum with spatial consensus to rectify boundary ambiguities. Extensive experiments on three benchmarks demonstrate that StrucNID achieves consistent and significant performance improvements across diverse datasets, validating the efficacy of our structural disentanglement paradigm.**

***Index Terms*—New Intent Discovery, Disentangled Representation Learning, Structural Graph Prior, large language models, Dialogue Systems.**

I. Introduction

New Intent Discovery (NID) [1], [2] is a pivotal capability for task-oriented dialogue systems, aiming to categorize known intents while simultaneously identifying novel user goals from unlabeled data streams. This task underpins the adaptability of modern AI Agents, ranging from customer service chatbots [3] to intelligent search engines, enabling them to recognize emerging user needs without extensive human annotation or retraining.

Existing embedding-centric methods [4]–[6] typically rely on deep clustering pipelines that encode utterances into monolithic latent representations. However, this paradigm inevitably suffers from Semantic Entanglement, where core intent semantics are inextricably confounded by instance-specific variations, such as diverse phrasing, tones, or non-semantic tokens [7], [8]. This entanglement causes samples sharing the same intent to scatter across the feature space rather than forming compact clusters, blurring decision boundaries and degrading performance, particularly in fine-grained scenarios.

*Corresponding author: Jindian Su (sujd@scut.edu.cn)

Conversely, recent LLM-Enhanced works [9]–[11] attempt to harness generalized knowledge but primarily utilize LLMs for generative augmentation or independent tagging. While beneficial, these approaches often neglect structural guidance, treating samples in isolation without reliable topological constraints. Consequently, they face Boundary Ambiguity, where outlier noise is frequently misclassified as novel intents due to the absence of a global structural prior to guide the clustering process.

To address these challenges, we present StrucNID, a robust framework that synergizes a Knowledge-Guided Structural Prior with Orthogonal Disentanglement. By leveraging structural topology to guide semantic learning, our approach effectively isolates core intents from stylistic noise and rectifies boundary uncertainties, ensuring robust discovery in open-world scenarios. Our main contributions are summarized as follows:

- We leverage LLMs to construct Semantic Mini-Clusters and model them as a high-precision graph, providing robust topological constraints for cold-start initialization rather than simple data augmentation.
- We project representations into orthogonal intent and instance subspaces, effectively filtering stylistic noise to purify clustering semantics.
- We design an adaptive threshold partitioning strategy to identify boundary samples, then apply Spatio-Temporal voting to dynamically rectify their ambiguities by integrating temporal momentum with spatial neighborhood consensus.
- Extensive experiments on three benchmarks demonstrate that StrucNID significantly outperforms competitive baselines, validating the efficacy of our structural disentanglement framework.

II. Related Work

A. New Intent Discovery

New Intent Discovery (NID) aims to identify known intents and group unknown ones from unlabeled data. Early works like DeepAligned [1] and CDAC+ [6] adopt iterative pseudo-labeling strategies, while recent methods like USNID [2] and MTP-CLNN [4] leverage contrastive learning to improve representations. However, these methods encode entire utterances into monolithic vectors, implicitly assuming geometric proximity correlates with semantic similarity. This assumption often fails when stylistic variations (e.g., phrasing, tone) dominate the distance metric, causing intra-class scatter [12]. These holistic approaches cannot effectively separate core intent from instance-specific noise, leading to degraded performance in low-resource scenarios.

B. LLM-Enhanced Intent Discovery

Recently, Large Language Models (LLMs) have been leveraged to enhance NID. Methods like IntentGPT [11] and IDAS [13] utilize LLMs to generate synthetic data or summaries, while ALUP [9] and LANID [10] employ LLM feedback to refine pseudo-labels. Despite their effectiveness, these approaches primarily use LLMs for data augmentation or iterative refinement, without providing structural guidance for representation learning [3]. This may introduce hallucinated noise that propagates through training. In contrast, StrucNID adopts a one-off structural construction strategy, leveraging LLMs solely to build a topological prior during initialization rather than iterative generation. While LLM-generated tags may introduce noise, our binary adjacency design eliminates false positives during cold-start, and the subsequent representation learning with Adaptive Consensus can dynamically rectify misclassified samples, mitigating the impact of initial label imperfections.

C. Disentangled Representation Learning

Disentangled Representation Learning (DRL) aims to identify and factorize the underlying generative factors hidden in data, a paradigm extensively applied in Computer Vision to decouple invariant identity from varying style features [14], [15]. However, this concept remains underexplored in NID. Standard models typically entangle intent semantics with instance-specific stylistic patterns within a unified feature space, leading to overfitting on superficial shortcuts rather than robust intent features. To address this, StrucNID introduces Orthogonal Disentanglement, mathematically constraining the model to separate the latent space into distinct intent and instance-specific subspaces. This yields cleaner, noise-resilient representations by effectively filtering out stylistic noise.

III. METHODOLOGY

As illustrated in Fig. 1, StrucNID integrates a Knowledge-Guided Structural Prior for initialization and an Intent-Instance Disentanglement mechanism for noise filtering, complemented by a Distribution-Aware Adaptive Consensus for boundary refinement.

A. Problem Formulation

We follow the standard semi-supervised NID setting [1], [2]. Unlike purely unsupervised clustering, we leverage prior knowledge from a labeled set $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^{N_l}$ with known intents $\mathcal{Y}_{known}$. The larger unlabeled set $\mathcal{D}^u = \{x_i\}_{i=1}^{N_u}$ contains utterances from both $\mathcal{Y}_{known}$ and a set of disjoint novel intents $\mathcal{Y}_{new}$ ($\mathcal{Y}_{known} \cap \mathcal{Y}_{new} = \emptyset$). The objective is to utilize $\mathcal{D}^l$ to simultaneously classify known classes and discover novel clusters in $\mathcal{D}^u$.

Adaptive Layer-wise Aggregation. Standard methods typically rely on the final [CLS] token, which may be overly specialized to pre-training objectives. To mitigate such entanglement, we employ a dynamic aggregation strategy that selectively combines hidden states from the last $M = 4$ layers (empirically $M = 4$), balancing semantic richness and computational efficiency. Let $H_i^{(l)} \in \mathbb{R}^{L \times d}$ denote the hidden states of the l-th layer, where L is the sequence length and $d = 768$ is the BERT-base hidden dimension. The aggregated representation $h_i \in \mathbb{R}^d$ is computed via a learnable weighted sum:

$$h_i = \text{LayerNorm}\left(\sum_{m=0}^{M-1} w_m \cdot \text{MeanPool}(H_i^{(N-m)})\right) \quad (1)$$

where $N = 12$ is the total number of BERT-base [16] layers and w_m are learnable scalar weights.

Adaptive K-Estimation. To estimate the number of clusters K, we adopt the Global Density Filtering strategy [17]. We utilize a high-density threshold $\delta_{th} \in [0.5, 0.8]$ to filter out low-confidence outliers from an initial over-clustering result, setting K as the count of remaining dense clusters.

B. Knowledge-Guided Structural Prior

Relying solely on geometric distances of raw BERT embeddings is unreliable for initialization due to the lack of domain-specific adaptation. To address this, we harness the semantic reasoning capability of LLMs to construct a Structural Prior.

Prompt-Conditioned Graph Construction. We treat intent discovery as a generative tagging task. For each utterance x_i, we construct a context C_i containing the raw text and a task-specific instruction (e.g., *"Extract a concise intent tag"*). We prompt an LLM to generate a fine-grained semantic tag S_i (a text phrase) based solely on the utterance content:

$$S_i = \arg\max_{w \in \mathcal{V}^*} P_{\Theta_{\text{LLM}}}(w \mid C_i) \quad (2)$$

For instance, given utterances "I want to transfer money" and "How do I send funds?", the LLM generates S_i="transfer money". We then map S_i to a discrete micro-label $y_i^{mini} = \phi(S_i)$ via canonicalization (lowercasing and stopword removal), e.g., ϕ("Transfer the Money")="transfer_money". To handle paraphrases and LLM variations, semantically equivalent tags (e.g., "transfer_money" and "send_funds") are merged into a unified identifier via fuzzy matching (selecting the most frequent as representative), assigning them the same y^{mini} to form Semantic Mini-Clusters. This process leverages the LLM's general semantic understanding without label supervision to provide initial grouping for subsequent representation learning.

Crucially, we materialize this prior knowledge into a global binary adjacency matrix $A^{(0)} \in \{0,1\}^{N \times N}$ that encodes pairwise mini-cluster membership, as visualized in Fig. 2:

$$A_{ij}^{(0)} = \begin{cases} 1, & \text{if } y_i^{mini} = y_j^{mini} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Rationale for Hard Adjacency. Binary adjacency maximizes precision during cold-start initialization, eliminating false positives from soft weighting that could propagate errors.

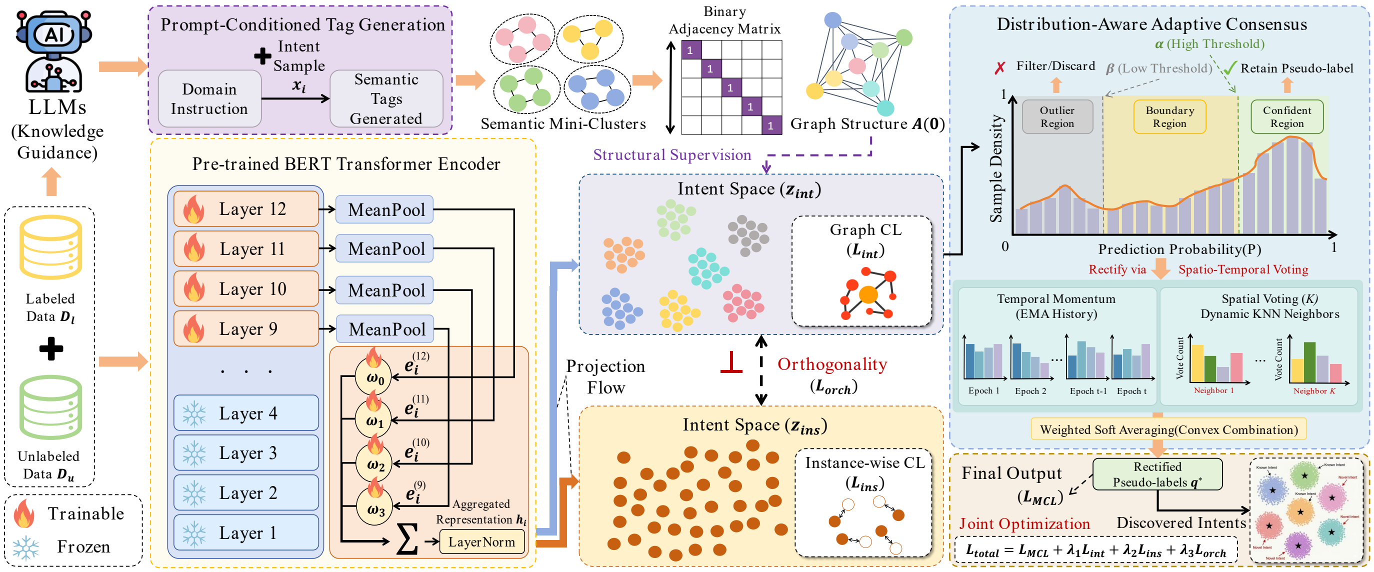
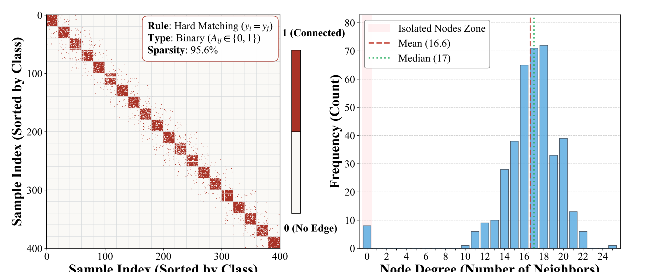


Fig. 1. Overview of our StrucNID framework. The framework leverages an LLM-guided structural prior for topological initialization. It then employs a dual-stream mechanism with orthogonal constraints to achieve intent-instance disentanglement for noise filtering. Finally, a distribution-aware consensus module rectifies boundary samples via Spatio-Temporal voting for robust clustering.



(a) Binary Semantic Graph Structure ($A^{(0)}$) (b) Node Degree Distribution (Connectivity)

Fig. 2. Visualization of the Knowledge-Guided Structural Prior.

As shown in Fig. 2(b), the connected components form reliable topological anchors, establishing a robust structural skeleton that prevents early-stage noise propagation. Isolated nodes are handled by subsequent Adaptive Consensus.

C. Intent-Instance Disentanglement

To rigorously separate class-invariant semantics from instance-specific noise, we propose a dual-stream projection mechanism. Unlike CV tasks where objects and backgrounds are visually distinct, *Semantic Entanglement* in NLP is more subtle, as intent and style are often interwoven in the same tokens.

Dual-Space Projection. The shared representation $h_i \in \mathbb{R}^d$ is projected into two orthogonal subspaces via independent non-linear heads:

$$z_i^{int} = g_{int}(h_i), \quad z_i^{ins} = g_{ins}(h_i) \tag{4}$$

where $z_i^{int} \in \mathbb{R}^{d'}$ and $z_i^{ins} \in \mathbb{R}^{d'}$ are the intent and instance representations respectively (typically $d' = 128$). Here, z_i^{int} is optimized to capture the Intent Semantics, while z_i^{ins} is designed to absorb the residual instance-specific variations (e.g., tone, phrasing).

Graph-Aware Contrastive Learning. We utilize the structural prior $A^{(0)}$ to strictly supervise the Intent Space. We apply Graph Contrastive Learning to pull together samples within the same reliable mini-cluster:

$$\mathcal{L}_{int}^{(i)} = -\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\log\frac{\exp(\mathrm{sim}(z_i^{int}, z_j^{int})/\tau)}{\sum_{k\neq i}\exp(\mathrm{sim}(z_i^{int}, z_k^{int})/\tau)} \tag{5}$$

where $\mathcal{N}_i = \{j \mid A_{ij}^{(0)} = 1\}$ denotes the set of reliable neighbors, and $\tau > 0$ is the temperature coefficient controlling distribution concentration. This loss anchors the intent features to the LLM-derived topology.

Instance-wise Contrastive Learning. Simultaneously, to prevent feature collapse in the Instance Space, we apply standard instance-wise contrastive learning. We generate an augmented view $\tilde{z}_i^{ins}$ (via standard dropout or cut-off) and treat it as the positive pair for z_i^{ins}, while all other samples in the batch are negatives:

$$\mathcal{L}_{ins}^{(i)} = -\log\frac{\exp(\mathrm{sim}(z_i^{ins}, \tilde{z}_i^{ins})/\tau)}{\sum_{k\neq i}\exp(\mathrm{sim}(z_i^{ins}, z_k^{ins})/\tau)} \tag{6}$$

where τ is the same temperature coefficient as in Eq. 5. This objective forces z^{ins} to distinguish each utterance instance, ensuring that unique stylistic details are preserved in this stream rather than being discarded.

Orthogonality Constraint. To ensure strict disentanglement, we minimize the Frobenius norm of the cross-correlation matrix between the two subspaces. Let $\hat{Z}^{int} \in \mathbb{R}^{N\times d}$ and $\hat{Z}^{ins} \in \mathbb{R}^{N\times d}$ denote the batch-level feature matrices stacking all $\{z_i^{int}\}$ and $\{z_i^{ins}\}$ respectively. The orthogonality loss is defined as:

$$\mathcal{L}_{orth} = \left\|(\hat{Z}^{int})^\top \hat{Z}^{ins}\right\|_F^2 \tag{7}$$

Mechanism Analysis. While orthogonality constraints are established in other domains, their application here relies on a novel synergy with the Structural Prior. Without $\mathcal{L}_{int}$, minimizing $\mathcal{L}_{orth}$ could lead to trivial solutions. In StrucNID, since z^{int} is anchored to semantic intent via $\mathcal{L}_{int}$ and z^{ins} captures instance identity via $\mathcal{L}_{ins}$, the orthogonality constraint effectively compels z^{ins} to capture only the information uncorrelated with intent, namely the stylistic noise.

D. Distribution-Aware Adaptive Consensus

To resolve boundary ambiguity, we rectify pseudo-labels using dynamic statistics.

Adaptive Region Partition. At epoch t, we set thresholds α_t and β_t as the 75^{th} and 25^{th} percentiles of the maximum prediction probabilities $p_{i,max}$. We formally define the region partition function $R(x_i)$ as:

$$R(x_i) = \begin{cases} \mathcal{P} \text{ (Confident)}, & \text{if } p_{i,max} > \alpha_t \\ \mathcal{B} \text{ (Boundary)}, & \text{if } \beta_t \leq p_{i,max} \leq \alpha_t \\ \mathcal{N} \text{ (Outlier)}, & \text{if } p_{i,max} < \beta_t \end{cases} \tag{8}$$

Spatio-Temporal Rectification. For boundary samples $x_i \in \mathcal{B}$, relying solely on instant predictions proves unstable. We therefore rectify their targets $q_i^* \in \mathbb{R}^K$ by aggregating evidence from two dimensions: Temporal Momentum and Spatial Consensus. Specifically, we enforce temporal consistency by maintaining an Exponential Moving Average (EMA) of historical predictions $\bar{p}_i^{(t)} \in \mathbb{R}^K$, updated as

$$\bar{p}_i^{(t)} = \mu \bar{p}_i^{(t-1)} + (1-\mu) P(y|z_i^{int}) \tag{9}$$

where $P(y|z_i^{int}) \in \mathbb{R}^K$ denotes the predicted probability distribution, and $\mu \in [0.8, 0.95]$ controls historical weight. Simultaneously, we retrieve the K-nearest neighbors $\mathcal{K}_i$ (typically $K = 20$) and compute a soft vote by averaging their distributions. The final rectified target is a convex combination with $\gamma \in [0.3, 0.7]$:

$$q_i^* = \gamma \bar{p}_i^{(t)} + (1-\gamma) \sum_{j \in \mathcal{K}_i} \frac{1}{|\mathcal{K}_i|} P(y|z_j^{int}) \tag{10}$$

By integrating historical stability and local neighborhood evidence, this mechanism effectively corrects ambiguous predictions at the decision boundaries.

E. Joint Optimization

We employ a Masked Clustering Loss ($\mathcal{L}_{MCL}$) that supervises the model only on confident and rectified boundary samples ($\mathcal{P} \cup \mathcal{B}$):

$$\mathcal{L}_{MCL} = -\frac{1}{|\mathcal{P} \cup \mathcal{B}|} \sum_{i \in \mathcal{P} \cup \mathcal{B}} \sum_{k=1}^{K} q_{i,k}^* \log P(k \mid z_i^{int}) \tag{11}$$

The total objective function is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{MCL} + \lambda_1 \mathcal{L}_{int} + \lambda_2 \mathcal{L}_{ins} + \lambda_3 \mathcal{L}_{orth} \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters balancing the contribution of each term. By minimizing $\mathcal{L}_{total}$, StrucNID synergistically optimizes feature disentanglement and robust cluster assignment.

IV. Experiments

A. Experimental Setup

Datasets. We conduct experiments on three benchmark datasets with varying granularity. The detailed statistics of each dataset is shown in Table I.

- **BANKING** [18]: A fine-grained dataset with 77 intents in the banking domain (13,083 utterances).
- **CLINC** [19]: A coarse-grained, multi-domain dataset with 150 intents across 10 domains (22,500 utterances).
- **StackOverflow** [20]: A technical dataset with 20 intents (20,000 utterances).

TABLE I
Statistics for employed benchmarks.

Dataset	#Utterances	#Intents	Length (max/mean)	Domain
BANKING	13,083	77	79/11.91	banking
CLINC	22,500	150	28/8.31	general
StackOverflow	20,000	20	41/9.18	questions

Baselines. We compare StrucNID with a comprehensive set of state-of-the-art methods, categorized into two groups:

- **Unsupervised Baselines.** We include representative deep clustering methods that learn representations solely from unlabeled data without using any known intent labels: **DCN** [5], **SAE-KM** [21], **DeepCluster** [17] and **USNID** [2].
- **Semi-supervised Baselines.** We compare against strong NID frameworks that leverage prior knowledge from known classes: **DWGF** [12], **MTP-CLNN** [4], and **CsePL** [22]. Additionally, we evaluate **USNID** [2] in its semi-supervised setting. To validate structural priors against generative augmentation, we also compare with LLM-enhanced methods: **ALUP** [9], **IntentGPT** [11], and **LANID** [10].

Evaluation Metrics. We adopt three standard clustering metrics: Accuracy (ACC) based on the Hungarian algorithm, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI). Higher values indicate better performance. Table II reports the performance on three benchmarks. For evaluation, we set the Known Class Ratio (KCR) to 0.75, and the labeled ratio of known intent classes to 0.1, following [1].

Implementation Details. We adopt BERT-base [16] as the backbone encoder. And we employ the AdamW optimizer with a learning rate of $2e^{-5}$ for the encoder and $1e^{-4}$ for the projection heads, with a batch size of 128. The temperature τ is set to 0.07 in Eq. 5, and loss weights are empirically set to $\lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 0.1$ in Eq. 12. For the structural prior construction, we utilize `GPT-4o-Mini` as the LLM. All experiments are conducted on 2 NVIDIA RTX 3090 GPUs, and results are averaged over 10 random seeds.

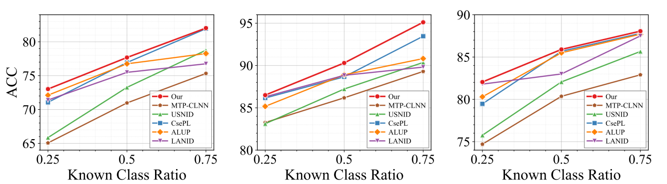
B. Main Results

Overall Performance Analysis. As evidenced in Table II, StrucNID consistently outperforms both embedding-based methods (e.g., MTP-CLNN) and recent LLM-enhanced

TABLE II
COMPARISON AGAINST THE UNSUPERVISED (UNSUP.) AND SEMI-SUPERVISED (SEMI-SUP.) STATE-OF-THE-ART ON THREE DATASETS. METHODS MARKED WITH † DENOTE LLM-ENHANCED APPROACHES. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND ARE <u>UNDERLINED</u>.

	Methods	BANKING			CLINC			StackOverflow			Average		
		NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC
Unsup.	DCN	62.72	25.36	38.59	74.77	31.68	48.69	58.75	36.23	59.48	65.41	31.09	48.92
	SAE-KM	59.80	23.59	37.07	73.77	31.58	47.15	44.96	28.23	49.11	59.51	27.80	44.44
	DeepCluster	41.77	8.95	20.69	65.58	19.11	35.70	17.52	3.09	18.64	41.62	10.38	25.01
	USNID	75.30	43.33	54.83	91.00	68.54	75.87	72.00	52.25	69.28	79.43	54.71	66.66
Semi-sup.	DWGF	86.41	68.16	79.38	**96.89**	<u>90.05</u>	<u>94.49</u>	81.73	75.30	87.60	88.34	77.84	87.16
	MTP-CLNN	84.63	64.32	75.33	95.44	84.23	89.30	73.88	64.04	79.68	84.65	70.86	81.44
	USNID	87.41	69.54	78.36	96.42	86.77	90.36	80.13	74.90	85.66	87.99	77.07	84.79
	CsePL	<u>87.70</u>	<u>71.36</u>	<u>81.93</u>	96.58	88.88	93.46	<u>82.81</u>	**75.99**	<u>87.80</u>	<u>89.03</u>	<u>78.74</u>	<u>87.73</u>
	ALUP†	86.58	70.33	78.27	95.80	85.23	89.69	81.21	72.63	85.40	87.86	76.06	84.45
	IntentGPT†	85.94	66.66	77.21	96.06	84.76	88.76	-	-	-	-	-	-
	LANID†	87.18	68.56	76.75	96.08	85.25	89.81	75.30	64.71	77.42	86.19	72.84	81.33
	StrucNID(ours)	**88.10**	**71.58**	**82.03**	<u>96.65</u>	**90.12**	**95.11**	**82.95**	<u>75.84</u>	**88.06**	**89.23**	**79.18**	**88.40**

baselines (e.g., ALUP) on average metrics. While CsePL achieves competitive performance via semantic-aware contrastive learning, StrucNID achieves further improvements across all datasets (average ACC: 88.40% vs. 87.73%). Notably, on the challenging fine-grained BANKING dataset, our method achieves substantial gains, surpassing IntentGPT by +4.82% and LANID by +5.28% in accuracy. Unlike LLM-enhanced approaches that primarily leverage LLMs for data augmentation, or embedding methods that suffer from semantic entanglement, StrucNID explicitly addresses the intrinsic interference between intent semantics and stylistic noise. This superior robustness confirms that our Intent-Instance Disentanglement mechanism offers a more effective solution for clarifying decision boundaries, particularly in scenarios with subtle semantic nuances.



(a) Effect on BANKING (b) Effect on CLINC (c) Effect on StackOverflow

Fig. 3. The effect of Known Class Ratio (KCR) on three benchmarks.

C. Impact of Known Class Ratio

Fig. 3 illustrates performance trends under varying Known Class Ratios (KCR $\in \{25\%, 50\%, 75\%\}$). While all methods benefit from increased supervision, StrucNID consistently outperforms competitive baselines across all ratios. Notably, even in the low-resource scenario (KCR=25%), our method maintains significant performance lead, demonstrating superior robustness in transferring semantic knowledge from limited known intents to discover novel categories.

D. Ablation Study

To verify the contribution of each module, we conduct an ablation study on BANKING shown in Table III.

TABLE III
ABLATION STUDY ON BANKING (KCR=75%). VALUES IN RED INDICATE PERFORMANCE DROPS COMPARED TO THE FULL MODEL.

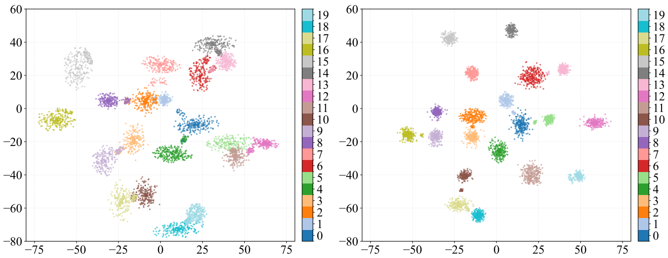
Methods	ACC	NMI	ARI
StrucNID (Full Model)	**82.03**	**88.10**	**71.58**
- w/o Structural Prior ($A^{(0)}$)	77.65$_{(\downarrow 4.38)}$	84.55$_{(\downarrow 3.55)}$	65.12$_{(\downarrow 6.46)}$
- w/o Disentanglement ($\mathcal{L}_{orth}$)	79.40$_{(\downarrow 2.63)}$	86.45$_{(\downarrow 1.65)}$	67.85$_{(\downarrow 3.73)}$
- w/o Adaptive Consensus	80.95$_{(\downarrow 1.08)}$	87.35$_{(\downarrow 0.75)}$	69.90$_{(\downarrow 1.68)}$
- w/o Layer Aggregation (h_i)	81.55$_{(\downarrow 0.48)}$	87.80$_{(\downarrow 0.30)}$	71.05$_{(\downarrow 0.53)}$

As observed, removing any component degrades performance. Specifically, ablating the structural prior $A^{(0)}$ causes the largest drop (ACC $\downarrow$4.38%), confirming its critical role in cold-start initialization. Excluding orthogonal disentanglement ($\mathcal{L}_{orth}$) incurs a 2.63% decline, validating its effectiveness in filtering stylistic noise. Removing adaptive consensus results in a 1.08% decrease, underscoring its necessity for boundary sample rectification. Finally, ablating layer aggregation leads to a 0.48% drop, demonstrating that hierarchical features capture richer semantics than single-layer embeddings.

E. Qualitative Analysis

Visualization of Representations. Fig. 4 visualizes the t-SNE embeddings of the test set. Compared to the baseline USNID [2], which shows blurred boundaries between similar intents (e.g., *card_payment* vs. *transfer*), StrucNID generates significantly more compact and well-separated clusters, validating the efficacy of our disentanglement strategy.

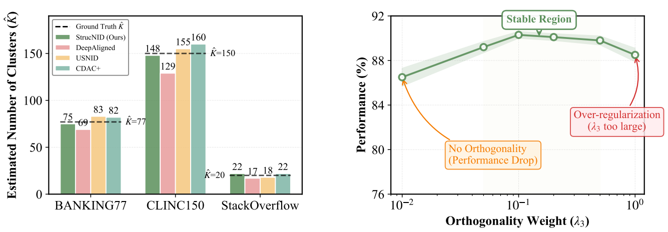
Estimation of Novel Intents (K). We evaluate the accuracy of estimating the number of categories. As shown in Fig. 5(a), StrucNID achieves remarkable precision across all datasets. Specifically on CLINC150 (Ground Truth $K = 150$), StrucNID estimates $\hat{K} = 148$ (Error: 1.3%), whereas baselines typically overestimate. This demonstrates that our Distribution-Aware Adaptive Consensus effectively identifies outlier noise, preventing it from forming false clusters.



(a) Strong Baseline **(b) StrucNID (Ours)**

Fig. 4. The t-SNE visualizations of learned intent representations on the StackOverflow dataset (KCR=75%).

Hyperparameter Sensitivity. We analyze the impact of the orthogonality weight λ_3 on the BANKING dataset at KCR=75% in Fig. 5(b). The performance peaks at $\lambda_3 = 0.1$ and remains consistently stable within the range $\lambda_3 \in [0.05, 0.5]$, indicating that StrucNID is robust to parameter selection and requires minimal tuning in practical deployment.



(a) K-Estimation Accuracy **(b) Hyperparameter Sensitivity**

Fig. 5. Robustness Analysis.

V. Conclusion

In this paper, we presented StrucNID, a robust framework addressing semantic entanglement in New Intent Discovery. By leveraging LLM-Guided structural priors and enforcing orthogonal disentanglement, we effectively filter out stylistic noise to form compact intent clusters while avoiding generation hallucinations that plague existing LLM-Enhanced methods. Our Spatio-Temporal voting mechanism further rectifies boundary ambiguities, reducing error propagation in open-world scenarios. Experimental results indicate substantial improvements in accuracy and stability across diverse benchmarks, particularly in fine-grained intent recognition tasks. Future work will explore multi-turn dialogue scenarios and multilingual intent discovery for real-world application.

References

[1] H. Zhang, H. Xu, T.-E. Lin, and R. Lyu, "Discovering new intents with deep aligned clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 365–14 373.

[2] H. Zhang, H. Xu, X. Wang, F. Long, and K. Gao, "A clustering framework for unsupervised and semi-supervised new intent discovery," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 5468–5481, 2024.

[3] I.-F. Lin, F. Hasibi, and S. Verberne, "SPILL: Domain-adaptive intent clustering based on selection and pooling with large language models," in *Findings of the Association for Computational Linguistics: ACL*, 2025, pp. 15 723–15 737.

[4] Y. Zhang, H. Zhang, L.-M. Zhan, X.-M. Wu, and A. Lam, "New intent discovery with pre-training and contrastive learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 256–269.

[5] X. Shen, Y. Sun, Y. Zhang, and M. Najmabadi, "Semi-supervised intent discovery with contrastive learning," in *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, 2021, pp. 120–129.

[6] T.-E. Lin, H. Xu, and H. Zhang, "Discovering new intents via constrained deep adaptive clustering with cluster refinement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8360–8367.

[7] X. Zhang, Q. Zhou, H. Liu, and H. Yu, "Pairwise prompt-based tuning with parameter efficient fast adaptation for generalized zero-shot intent detection," in *Findings of the Association for Computational Linguistics: NAACL*, 2025, pp. 917–929.

[8] Y. Zhou, G. Quan, and X. Qiu, "A probabilistic framework for discovering new intents," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 3771–3784.

[9] J. Liang, L. Liao, H. Fei, B. Li, and J. Jiang, "Actively learn from LLMs with uncertainty propagation for generalized category discovery," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024, pp. 7845–7858.

[10] L. Fan, J. Pu, R. Zhang, and X.-M. Wu, "LANID: LLM-assisted new intent discovery," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024.

[11] J. A. Rodriguez, N. Botzer, D. Vazquez, C. Pal, M. Pedersoli, and I. Laradji, "Intentgpt: Few-shot intent discovery with large language models," *arXiv preprint arXiv:2411.10670*, 2024.

[12] W. Shi, W. An, F. Tian, Q. Zheng, Q. Wang, and P. Chen, "A diffusion weighted graph framework for new intent discovery," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 8033–8042.

[13] M. De Raedt, F. Godin, T. Demeester, and C. Develder, "IDAS: Intent discovery with abstractive summarization," in *Proceedings of the 5th Workshop on NLP for Conversational AI*, 2023, pp. 71–88.

[14] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9677–9696, 2024.

[15] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1283–1292.

[16] 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*, 2019, pp. 4171–4186.

[17] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Computer Vision – ECCV 2018: 15th European Conference*, 2018, p. 139–156.

[18] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić, "Efficient intent detection with dual sentence encoders," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 2020, pp. 38–45.

[19] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars, "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)*, 2019, pp. 1311–1316.

[20] J. Xu, P. Wang, G. Tian, B. Xu, J. Zhao, F. Wang, and H. Hao, "Short text clustering via convolutional neural networks," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015, pp. 62–69.

[21] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 2016, p. 478–487.

[22] J. Liang and L. Liao, "ClusterPrompt: Cluster semantic enhanced prompt learning for new intent discovery," in *Findings of the Association for Computational Linguistics: EMNLP*, 2023, pp. 10 468–10 481.