A Survey of Ontology Expansion for Conversational Understanding

Jinggui Liang[1], **Yuxia Wu**[1], **Yuan Fang**[1], **Hao Fei**[2], **Lizi Liao**[1]

[1]Singapore Management University, [2]National University of Singapore

jg.liang.2023@phdcs.smu.edu.sg   yieshah2017@gmail.com
yfang@smu.edu.sg   haofei37@nus.edu.sg   lzliao@smu.edu.sg

Abstract

In the rapidly evolving field of conversational AI, **On**tology **Exp**ansion (**OnExp**) is crucial for enhancing the adaptability and robustness of conversational agents. Traditional models rely on static, predefined ontologies, limiting their ability to handle new and unforeseen user needs. This survey paper provides a comprehensive review of the state-of-the-art techniques in OnExp for conversational understanding. It categorizes the existing literature into three main areas: (1) *New Intent Discovery*, (2) *New Slot-Value Discovery*, and (3) *Joint OnExp*. By examining the methodologies, benchmarks, and challenges associated with these areas, we highlight several emerging frontiers in OnExp to improve agent performance in real-world scenarios and discuss their corresponding challenges. This survey aspires to be a foundational reference for researchers and practitioners, promoting further exploration and innovation in this crucial domain.

1 Introduction

Conversational understanding (CU) is a core component in the development of conversational agents (Li et al., 2017; Carmel et al., 2018). The objective of the CU module is to accurately capture and interpret user needs during interactions. As illustrated in Figure 1, these capabilities are generally encapsulated within a conversational ontology, which defines a collection of possible user intents, slots, and values for each slot (Mrksic et al., 2017; Budzianowski et al., 2018; Neves Ribeiro et al., 2023). Effective CU models must not only identify the overall purposes (*intent detection*) (E et al., 2019) expressed by users but also pinpoint relevant pieces of information (*slot filling*) (Wang et al., 2021a) that fulfill these intents.

Traditionally, CU research assumes a well-defined, static ontology where all intents, slots, and most possible values are predetermined. Within
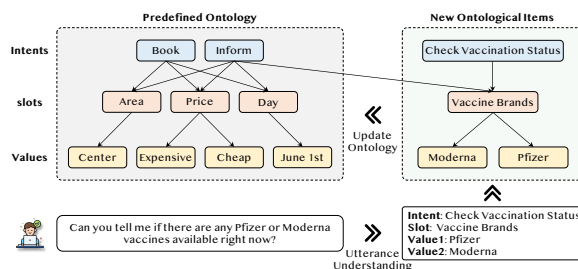

Figure 1: An example of ontology expansion enabling conversational agents to adapt to unseen events.

this predefined framework, CU is often treated as a closed-world classification task for intents and sequence labeling task for slot values (Larson and Leach, 2022). However, in real-world settings, conversational agents encounter rapidly evolving user needs and diverse expressions, leading to the emergence of new ontological items (Liang and Liao, 2023; An et al., 2024). This dynamic environment presents a significant challenge, as traditional CU models fail easily in situations beyond the predefined ontology.

To address this challenge, OnExp has been proposed to facilitate open-world ontology learning (Lin et al., 2020; Zhang et al., 2021c,b, 2022; Wu et al., 2022a). It dynamically updates and extends the conversational ontology by recognizing both pre-established and novel ontological items from user utterances. Effective OnExp approaches can significantly enhance the downstream decision-making and policy implementation of conversational agents, improving user satisfaction and service efficiency (Dao et al., 2023, 2024).

Recent years have witnessed substantial progress in developing innovative OnExp methodologies. However, the rapid advancements have left a gap in comprehensive reviews that summarize these efforts and discuss emerging trends. This paper aims to fill this gap by providing a thorough survey of OnExp research. We introduce the preliminaries of

18111

Dataset	Domain	#Samples	#Ontologies		Supported Tasks	
			#Intents	#Slots	NID	NSVD
BANKING77 (Casanueva et al., 2020)	Bank	13,083	77	-	✓	✗
CLINC150 (Larson et al., 2019)	Multi-domain	22,500	150	-	✓	✗
StackOverflow (Xu et al., 2015)	Question	20,000	20	-	✓	✗
CamRest (Wen et al., 2017)	Restaurant	2,744	2	4	✗	✓
Cambridge SLU (Henderson et al., 2012)	Restaurant	10,569	5	5	✗	✓
WOZ-attr (Eric et al., 2020)	Attraction	7,524	3	8	✗	✓
WOZ-hotel (Eric et al., 2020)	Hotel	14,435	3	9	✗	✓
ATIS (Hemphill et al., 1990)	Flight	4,978	17	79	✓	✓
SNIPS (Coucke et al., 2018)	Multi-domain	13,784	7	72	✓	✓
SGD (Rastogi et al., 2020)	Multi-domain	329,964	46	214	✓	✓

Table 1: Summary of popular datasets for OnExp. #Samples, #Intents, and #Slots represent the total number of utterances, intents, and slots, respectively.

OnExp, detailing task formulations, data resources, and evaluation protocols. Our novel taxonomy categorizes OnExp studies into three types: (1) New Intent Discovery (NID), (2) New Slot-Value Discovery (NSVD), and (3) Joint OnExp, offering comprehensive coverage of the field. Finally, we discuss promising research directions and associated challenges, motivating further exploration.

In summary, our contributions are as follows:(1) We present the first comprehensive survey on ontology expansion; (2) We categorize OnExp research into three branches: NID, NSVD, and Joint OnExp, providing a unified understanding of the literature; (3) We discuss emerging frontiers and challenges in OnExp, highlighting future research directions. Additionally, we maintain a GitHub repository[1] that organizes useful resources.

2 Preliminaries

2.1 Task Formulation

Ontology expansion in conversational understanding involves dynamically broadening the predefined ontology by recognizing both known and novel ontological items from user utterances. These items are structured as a collection of intents, slots, and corresponding slot values.

Formally, let $\mathcal{O}_k$ and $\mathcal{O}_u$ represent the sets of predefined and unknown ontological items, with $\mathcal{O}_u \cap \mathcal{O}_k = \varnothing$. The OnExp tasks consider a dataset $\mathcal{D}^{all}$ that is divided into two parts: a labeled dataset $\mathcal{D}^l$ and an unlabeled dataset $\mathcal{D}^u$. $\mathcal{D}^l = \{(\boldsymbol{x}_i, o_i) | o_i \in \mathcal{O}_k\}_{i=1}^{|\mathcal{D}^l|}$ consists of utterances paired with labels that belong to $\mathcal{O}_k$. Conversely,

$\mathcal{D}^u = \{\boldsymbol{x}_i | o_i \in \mathcal{O}_u \cup \mathcal{O}_k\}_{i=1}^{|\mathcal{D}^u|}$ includes utterances for which the labels are not available during the model learning, covering both $\mathcal{O}_k$ and $\mathcal{O}_u$.

Given an utterance $\boldsymbol{x}_i \in \mathcal{D}^{all}$, the overall objective of OnExp tasks is to optimize a mapping function f_θ^{OnExp}, parameterized by θ, to recognize its corresponding ontological items as follows:

$$f_\theta^{OnExp}(\boldsymbol{x}_i) \to (o_i^I, o_i^S, o_i^V, r), \quad (1)$$

where $(o_i^I, o_i^S, o_i^V) \in \mathcal{O}_k \cup \mathcal{O}_u$ denote the intent, slot, and value associated with $\boldsymbol{x}_i$. The term r refers to the relations among various ontological items, such as the intent *Check Vaccination Status* being associated with the slot *Vaccine Brands*, but not with the slot *Area*. As the focus of OnExp is on identifying and expanding fundamental concepts emerging from dynamic conversations, the relations among these items are typically overlooked in the existing literature.

As discussed in Section 1, OnExp encompasses various tasks. In the NID setting, the mapping function f_θ^{OnExp} predicts only o^I, discarding (o^S, o^V). In the NSVD setting, the focus shifts to uncovering (o^S, o^V), omitting intents o^I. In Joint OnExp, (o^I, o^S, o^V) are all retained, with the aim of leveraging shared knowledge across these tasks for more effective ontology learning.

2.2 Data Resources

High-quality annotated datasets are essential for developing OnExp methods. We summarize the commonly used data resources, with an overview of each dataset's domain, scale, annotated ontological items, and supported tasks in Table 1.

For **NID**, the most widely used datasets are BANKING77 (Casanueva et al., 2020), CLINC150

(Larson et al., 2019), and StackOverflow (Xu et al., 2015). For **NSVD**, prominent datasets include CamRest (Wen et al., 2017), Cambridge SLU (Henderson et al., 2012), WOZ-attr (Eric et al., 2020), WOZ-hotel (Eric et al., 2020), ATIS (Hemphill et al., 1990), SNIPS (Coucke et al., 2018), and SGD (Rastogi et al., 2020). Further details on these datasets are provided in Appendix A.1.

2.3 Evaluation Protocols

NID Metrics. The NID evaluation metrics include: (1) Accuracy (**ACC**), based on the Hungarian algorithm; (2) Adjusted Rand Index (**ARI**); and (3) Normalized Mutual Information (**NMI**).

NSVD Metrics. The performance of NSVD systems is evaluated using the following key metrics: (1) **Precision**, (2) **Recall**, and (3) **F1-score**. The F1-score, which is calculated based on slot value spans, is also referred to as **Span-F1**.

Other Metrics. Notably, these evaluation metrics are not confined to the corresponding settings described previously. Additionally, the OnExp models can also be evaluated by **Known Acc**, **Novel Acc**, and **H-score** (An et al., 2024). Thorough discussions and specific definitions of the above evaluation metrics are detailed in Appendix A.2.

3 Taxonomy of OnExp Research

This section presents the new taxonomy for OnExp as shown in Figure 2, comprising *New Intent Discovery* (§3.1), *New Slot-Value Discovery* (§3.2), and *Joint OnExp* (§3.3).

3.1 New Intent Discovery

We first explore the NID task in this section, which aims to simultaneously identify known and newly emerged user intents. Notably, NID operates at the utterance level, excelling in isolating distinct user intents but struggling with overlapping or ambiguous ones. To achieve effective NID, a variety of methodologies have been devised, as illustrated in Figure 2. We classify these NID studies into three categories based on the use of available labeled data: Unsupervised NID, Zero-shot NID, and Semi-supervised NID.

3.1.1 Unsupervised NID

Unsupervised NID aims to discover user intents without any labeled data, facing significant challenges in deriving effective intent patterns to group similar utterances. This section categorizes existing unsupervised NID efforts into three types based on their model designs: Rule-based, Statistical, and Neural Network-based (NN-based) Methods.

Rule-based Methods. Early efforts, such as those by Rose and Levinson (2004), collaborated with domain experts to develop a conceptual schema for user goals, adapting to new goal categories. Jansen et al. (2008) used a decision tree for intent analysis. However, maintaining these rule-based models proved challenging as the complexity of rules intensified across different domains.

Statistical Methods. Given the limitations inherent in rule-based systems, statistical methods emerged as a more robust and effective alternative. Typical clustering algorithms like K-Means (MacQueen et al., 1967) and Agglomerative Clustering (Gowda and Krishna, 1978) laid the groundwork. Later, Aiello et al. (2011) aggregated fine-grained intent-related missions to learn new search intents, while Cheung and Li (2012) used external knowledge bases for sequence clustering. Methods like Ren et al. (2014) utilized heterogeneous graphs for cross-source intent learning, and Hakkani-Tür et al. (2013) introduced Bayesian models leveraging clicked URLs as implicit supervision in clustering new intents, while Hakkani-Tür et al. (2015) explored the lexical semantic structure of user utterances with semantic parsers. Despite their robustness, these methods often struggled with high-dimensional data and complex semantics.

NN-based Methods. To address the limitations of statistical methods, deep neural models have been explored for more effective new intent learning, thanks to their superior learning capabilities and flexible parameters. Xie et al. (2016) proposed Deep Embedded Clustering (DEC), which iteratively refines intent clusters using an auxiliary target distribution. Yang et al. (2017) developed a Deep Clustering Network (DCN) that combines nonlinear dimensionality reduction with K-Means clustering to optimize utterance representations. Deep Adaptive Clustering (DAC) (Chang et al., 2017) reimagined intent discovery as a pairwise classification problem, employing a binary-constrained model to learn relationships between utterance pairs. DeepCluster (Caron et al., 2018) alternated between clustering utterances and refining their representations via cluster assignments. Further advancements include Supporting Clustering
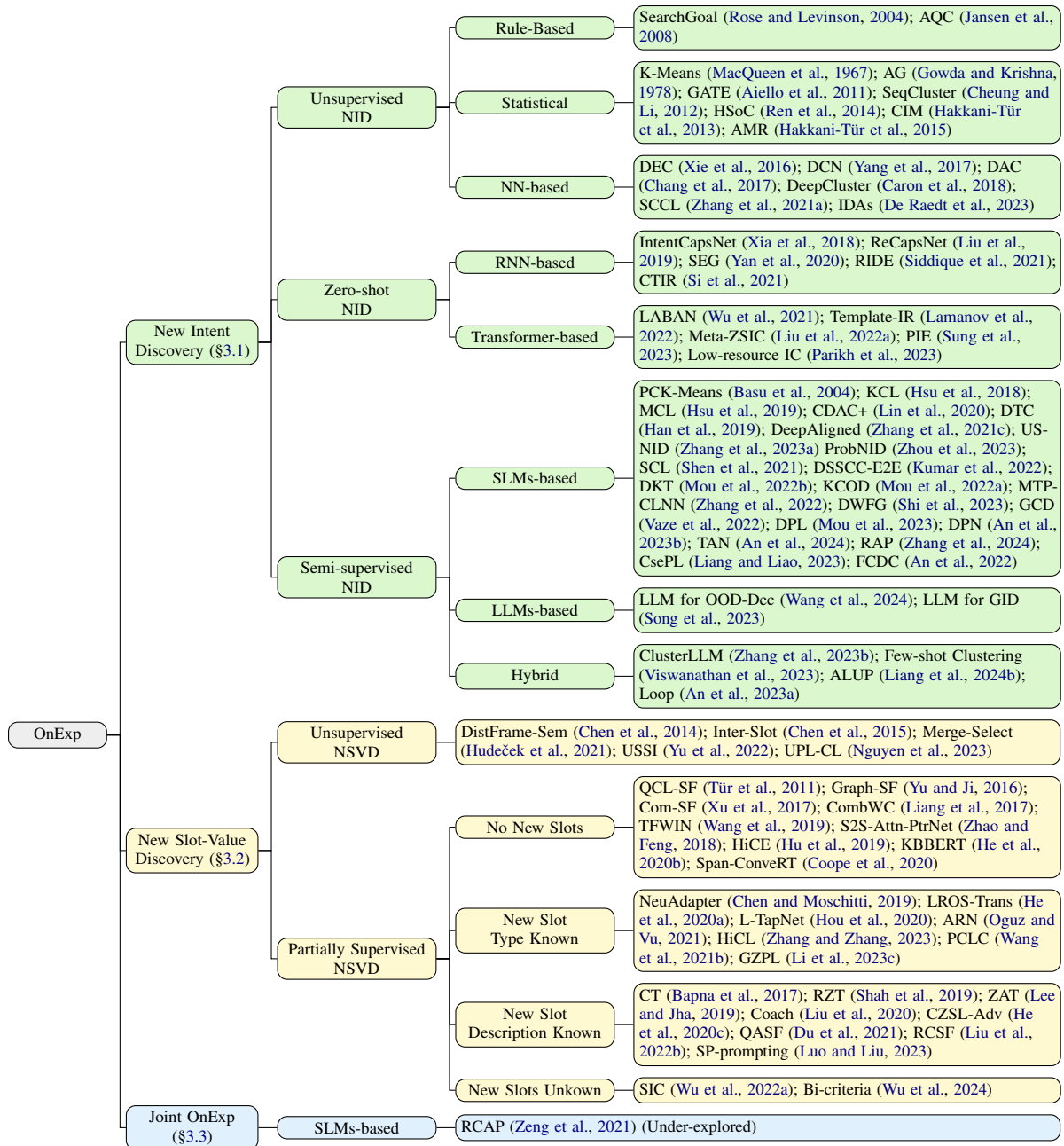
Figure 2: The taxonomy for Ontology Expansion.

with Contrastive Learning (SCCL) (Zhang et al., 2021a), which utilized emerging contrastive learning techniques to enhance intent clustering. In the era of Large Language Models (LLMs), (De Raedt et al., 2023; Liang et al., 2024a) further leveraged LLMs to enhance intent clustering.

3.1.2 Zero-shot NID.

Zero-shot NID aims to discover new user intents using only labeled training data from known intents. The main challenge lies in effectively transferring the prior knowledge of known intents to facilitate the recognition of new intents. This setting

is divided into RNN-based and Transformer-based methods based on their backbone architecture.

RNN-based Methods. RNNs were the dominant model for encoding sentences in the early days. Hence, Xia et al. (2018) proposed an RNN-based capsule network with routing-by-agreement to adapt the model to new intents. To address the polysemy problem, Liu et al. (2019) introduced a dimensional attention mechanism and learned generalizable transformation matrices for new intents. Beyond merely extracting features from utterances, Siddique et al. (2021) incorporated com-

monsense knowledge to learn robust relationship meta-features. Despite these advancements, Si et al. (2021) identified a critical issue: new intent representations cannot be learned during training. Hence, they proposed the Class-Transductive Intent Representations framework, which progressively optimizes new intent features using intent names.

Transformer-based Methods. In practice, the sequential nature of RNNs incurs high computational costs and struggles with long-range dependencies. To address these issues, Transformers have emerged as an effective solution for zero-shot NID. Wu et al. (2021) developed a label-aware BERT attention network that constructs an intent label semantic space to map utterances to intent labels. Following this, Lamanov et al. (2022) modeled this task as a sentence pair modeling problem, utilizing pre-trained language models to fuse intent labels and utterances for binary classification. Liu et al. (2022a) introduced a mixture attention mechanism and collaborated it with a novel meta-learning paradigm to enhance new intent identification. To better adapt pre-trained encoders to intent discovery, Sung et al. (2023) proposed generating pseudo-intent names from utterances and applied intent-aware contrastive learning to develop the Pre-trained Intent-aware Encoder (PIE). Recently, Parikh et al. (2023) explored zero-shot NID using Large Language Models (LLMs), investigating the various strategies such as in-context prompting to aid in identifying novel intents.

3.1.3 Semi-supervised NID.

Semi-supervised NID combines limited labeled data with extensive unlabeled data to discern new intents. This approach faces challenges in deriving supervision signals for unlabeled utterances and avoiding overfitting to known intents. Unlike Zero-shot NID, which is provided with new intent names or classes, semi-supervised NID does not know the new intents or their quantity. This section categorizes methods into Small Language Models (SLMs)-based, LLMs-based, and Hybrid methods.

SLMs-based Methods. SLMs like BERT, pre-trained on large-scale corpora, exhibit strong text understanding abilities and have been effectively fine-tuned for various tasks (Devlin et al., 2019; Lewis et al., 2020). Utilizing SLMs as feature extractors, Basu et al. (2004) introduced Pairwise Constrained K-Means (PCK-Means) with active constraint selection for new intent clustering.

Building on this, Hsu et al. (2018) used SLMs for static constraints with KL divergence-based Contrastive Loss (KCL), while Hsu et al. (2019) proposed Meta Classification Likelihood (MCL) for dynamic pairwise similarity updates. Lin et al. (2020) presented Constrained Deep Adaptive Clustering (CDAC+) for iterative model refinement.

Despite these advances, pairwise supervision signals often fall short in fully utilizing labeled data. To address this, Han et al. (2019) proposed Deep Transfer Clustering (DTC), improving clustering quality through consistency regulation and intent cluster number estimation. Zhang et al. (2021c) developed DeepAligned to resolve label inconsistencies, later improved by USNID for faster convergence (Zhang et al., 2023a). Zhou et al. (2023) alleviated prior knowledge forgetting with Prob-NID, a probabilistic framework optimizing intent assignments via Expectation Maximization. Zhang et al. (2022) utilized multi-task pre-training and K-nearest neighbor contrastive learning for compact clusters (MTPCLNN). Additionally, Shi et al. (2023) proposed the Diffusion Weighted Graph Framework (DWGF), capturing both semantic and structural relationships within utterances for more reliable supervisory signals. Beyond learning contrastive relations, An et al. (2023b) formulated a bipartite matching problem, proposing the Decoupled Prototypical Network (DPN) to separate known from new intents, facilitating explicit knowledge transfer. Zhang et al. (2024) introduced Robust and Adaptive Prototypical learning (RAP) to enhance intra-cluster compactness and inter-cluster dispersion. Recently, Liang and Liao (2023) leveraged prompt learning with two-level contrastive learning and soft prompting for new intent discovery.

While successful, SLM-based methods require extensive fine-tuning on large datasets, which is time-consuming. Moreover, SLMs struggle to fully capture the nuanced semantics of diverse and dynamic human languages in conversational contexts.

LLMs-based Methods. Recently, LLMs (OpenAI, 2023; Touvron et al., 2023) have shown impressive efficacy across a broad range of NLP tasks, such as summarization (Liu et al., 2023) and query rewriting (Anand et al., 2023; Guo et al., 2024). Given the above SLMs' limitations, there is a growing trend toward using LLMs for intent discovery in few/zero-shot settings. Wang et al. (2024) evaluated LLMs' ability to detect unknown intents, using ChatGPT to classify intents beyond the predefined

set. Moreover, Song et al. (2023) broadened the use of LLMs in intent discovery, directing ChatGPT to group utterances and identify known and novel intents.

Hybrid Methods. Although LLMs-based methods excel in zero-shot settings, they typically underperform compared to fully fine-tuned models. To address this, Hybrid methods that combine the strengths of SLMs and LLMs have been developed to enhance intent discovery. In this effort, Zhang et al. (2023b) proposed ClusterLLM, which uses triplet feedback from LLMs to refine SLMs-learned representations and applies pairwise hierarchical clustering to improve cluster granularity. Further, Viswanathan et al. (2023) investigated three strategies—keyphrase expansion, pairwise constraints, and cluster correction—to leverage LLMs for better intent clustering. To effectively utilize LLMs and reduce costs, Liang et al. (2024b) integrated LLMs into active learning, using uncertainty propagation to selectively label utterances and extending this feedback without spreading inaccuracies. Similarly, An et al. (2023a) introduced local inconsistent sampling with scalable queries to correct inaccurately allocated utterances using LLMs.

3.2   New Slot-Value Discovery

The NSVD task seeks to identify new slots and the corresponding values that emerge from dynamic conversations. Unlike the previous NID task that focuses on utterance-level recognition, NSVD specifically narrows its scope within individual utterances, excelling in detailed information extraction but limited by the quality and specificity of input data. Innovations in this task can be classified into unsupervised NSVD and partially supervised NSVD.

3.2.1   Unsupervised NSVD

Unsupervised NSVD discovers new slots and values without any labeled data, facing challenges such as dialogue noise and requiring high human intervention for ranking or selection processes. Early works like Chen et al. (2013) combined a frame-semantic parser with a spectral clustering-based slot ranking model to induce semantic slots. (Chen et al., 2014) further refined this method by integrating semantic frame parsing with word embeddings. Moreover, Chen et al. (2015) enhanced slot discovery by constructing lexical knowledge graphs and employing random walks to delineate slots. Despite the benefits of linguistic tools for discovering

new slots, such methods struggled with dialogue noise and the ranking processes require significant human intervention. Addressing these challenges, Hudeček et al. (2021) revised the ranking method to iteratively refine the obtained slots through slot taggers. To reduce reliance on generic parsers, Yu et al. (2022) further proposed a unified slot schema induction method that incorporates data-driven candidate value extraction and coarse-to-fine slot clustering. Recently, Nguyen et al. (2023) utilized pretrained language model probing combined with contrastive learning refinement to induce value segments for slot induction.

3.2.2   Partially Supervised NSVD

Partially supervised NSVD leverages some form of labeled data and is divided into four types based on the supervision nature: No New Slots, New Slot Type Known, New Slot Description Known, and New Slot Unknown.

No New Slots. This setting operates with all slot types predefined and certain known values for each slot labeled. It primarily explores leveraging existing slots to identify new values within these predefined slots, facing challenges in efficiently mining new value entities and leveraging external knowledge. This is common in scenarios where new restaurant names or new vaccine brand names emerge. Specifically, Tür et al. (2011) mined new slot entities from user queries in query click logs with target URLs, while Yu and Ji (2016) used dependency trees to identify slot-specific triggers. Xu et al. (2017) introduced a slot filler refinement method that constructs entity communities to filter out incorrect new fillers. Liang et al. (2017) combined word/character-level embeddings via highway networks to detect new values. Further, Wang et al. (2019) explored the temporal slot-filling problem and proposed a pattern-based framework that assesses pattern reliability and detects conflicts to find temporal values. To tackle the unknown value issue more effectively, Hu et al. (2019) formulated a K-shot regression problem, using a hierarchical context encoder and meta-learning to better infer new value embeddings. To explore the potential of external knowledge in aiding the discovery of new values, He et al. (2020b) employed background knowledge bases with a knowledge integration method to facilitate tagging slot values.

New Slot Type Known. Unlike merely identifying new values for predefined slots, practical ap-

Methods	BANKING77			CLINC150			StackOverflow		
	ACC	ARI	NMI	**ACC**	ARI	NMI	**ACC**	ARI	NMI
SLMs-based Methods									
PCK-Means (Basu et al., 2004)	32.66	16.24	48.22	54.61	35.40	68.70	24.16	5.35	17.26
BERT-KCL (Hsu et al., 2018)	60.15	46.72	75.21	68.86	58.79	86.82	13.94	7.81	8.84
BERT-MCL (Hsu et al., 2019)	61.14	47.43	75.68	69.66	59.92	87.72	72.07	57.43	66.81
CDAC+ (Lin et al., 2020)	53.83	40.97	72.25	69.89	54.33	86.65	73.48	52.59	69.84
BERT-DTC (Han et al., 2019)	56.51	44.70	76.55	74.15	65.02	90.54	71.47	53.66	63.17
DeepAligned (Zhang et al., 2021c)	64.90	53.64	79.56	86.49	79.75	93.89	-	-	-
MTPCLNN (Zhang et al., 2022)	73.98	63.10	84.22	88.25	84.77	94.88	83.18	69.50	77.03
ProbNID (Zhou et al., 2023)	74.03	62.92	84.02	88.99	83.00	95.01	80.50	65.70	77.32
DPN (An et al., 2023b)	74.45	63.26	84.31	89.22	84.30	95.14	84.59	70.27	79.89
RAP (Zhang et al., 2024)	76.27	65.79	85.16	91.24	86.28	95.93	86.60	71.73	82.36
USNID (Zhang et al., 2023a)	78.36	69.54	87.41	90.36	86.77	96.42	85.66	74.90	80.13
DFWG (Shi et al., 2023)	79.38	68.16	86.41	94.49	90.05	96.89	87.60	75.30	81.73
CsePL (Liang and Liao, 2023)	81.93	71.36	87.70	93.46	88.88	96.58	87.80	75.99	82.81
LLMs-based Methods									
LLM for GID (Song et al., 2023)	64.22	-	-	84.33	-	-	-	-	-
Hybrid Methods									
Few-shot Clustering (Viswanathan et al., 2023)	65.30	-	82.40	79.40	-	92.60	-	-	-
ClusterLLM (Zhang et al., 2023b)	71.20	-	85.15	83.80	-	94.00	-	-	-
ALUP (Liang et al., 2024b)	82.85	73.10	88.35	94.93	89.22	97.43	87.70	76.03	83.14

Table 2: The main semi-supervised NID results on three benchmarks.

plications may require models to extract values for well-defined slots not seen during training. The main challenge is adapting models to new slots. To address this, Chen and Moschitti (2019) explored transfer learning for labeling new values and developed a neural adapter to adapt previously trained models to these new slots. Further, He et al. (2020a) improved transfer learning efficiency by learning the label-relational output structure to capture slot label correlations, while Wang et al. (2021b) introduced prototypical contrastive learning with label confusion to refine slot prototypes dynamically. Beyond using coarse slot label information, (Zhang and Zhang, 2023) introduced Hierarchical Contrastive Learning (HiCL), where coarse and fine-grained slot labels serve as supervised signals to assist in extracting cross-domain slot fillers. Recently, Li et al. (2023c) explored advanced prompting techniques for identifying new values, using slot types and inverse prompting to enhance model performance.

New Slot Description Known. In contrast to accessing well-defined new slot types, this setting deals with extracting new values using only coarse-grained descriptions of new slots. Concretely, Bapna et al. (2017) proposed Concept Tagger (CT) for cross-domain slot-filling with slot descriptions,

while Shah et al. (2019) used slot descriptions to improve slot representations. In addition, Liu et al. (2020) proposed a coarse-to-fine (Coach) method that initially learns value patterns coarsely, then fills them into fine slot types based on the similarity with the representation of each slot type description. Inspired by this, He et al. (2020c) enhanced Coach with contrastive loss and adversarial attacks to improve robustness. Contrary to previous methods, Du et al. (2021) and Liu et al. (2022b) tackle the slot-filling problem as a reading comprehension task, extracting new values by answering questions derived from slot descriptions. Recently, Luo and Liu (2023) combined learnable prompt tokens and discrete tokens of slot descriptions to identify new values.

New Slot Unknown. Unlike the above studies, this setting focuses on extracting new slot values while also inducing potential new slots, without knowing the prior information of new slots. In this context, Wu et al. (2022a) used existing linguistic annotation tools to extract slot values and proposed an incremental clustering scheme that synergizes labeled and unlabeled data for slot structure discovery. To reduce labeling efforts with robust performance, Wu et al. (2024) introduced a Bi-criteria active learning scheme that selects data

Methods	CamRest	Cambridge SLU	WOZ-hotel	WOZ-attr	ATIS
CDAC+ (Lin et al., 2020)	20.4	17.8	17.4	55.2	58.2
BERT-DTC (Han et al., 2019)	13.1	13.8	17.0	54.5	54.3
DeepAligned (Zhang et al., 2021c)	66.3	63.3	37.8	64.4	62.9
SIC (Wu et al., 2022a)	70.6	77.0	58.8	76.1	63.8
Bi-criteria (Wu et al., 2024)	-	-	68.94	78.25	87.96

Table 3: The Span-F1 scores of New Slot Unknown methods on five benchmarks.

based on uncertainty and diversity when discerning new slots.

3.3 Joint OnExp

While significant successes have been achieved, previous methods tackle new intent and slot-value discovery as separate tasks, despite their inherent interconnection. Joint OnExp addresses this by simultaneously identifying new intents, slots, and values, offering a comprehensive understanding but posing challenges in managing knowledge sharing without compromising performance. Pioneers in this field, Zeng et al. (2021) devised a coarse-to-fine three-step method—role-labeling, concept-mining, and pattern-mining—to infer intents, slots, and values. Despite its promising results, Joint OnExp is still under-explored, offering substantial space for further innovation.

4 Leaderboard and Takeaway

Leaderboard: The leaderboard for representative NID and NSVD methods on widely recognized datasets is presented in Table 2 and Table 3. More details are presented in Appendix B.

Takeaway for NID: Based on the review of NID efforts, we present the following observations:

- *Pre-trained Language Models Enhance OnExp.* It has been observed that NID methods utilizing pre-trained models, such as CsePL and ALUP, consistently outperform traditional methods like PCK-Means by significant margins ($\sim 50\%$ in ACC). This demonstrates that pre-trained models, including LLMs, contribute substantial foundational knowledge and supplementary supervision signals. They enhance NID performance by offering a deeper contextual understanding and quicker adaptation to new user intents.

- *Prior Knowledge Leads to Improvement.* We observe that NID methods with supervision generally surpass unsupervised ones, as incorporating prior knowledge—through labeled data or

external information—significantly boosts the model's ability to identify new intents. For example, semi-supervised CsePL shows over 5% improvements in all evaluation metrics compared to the SOTA unsupervised IDAS. This highlights the critical role of integrating prior knowledge.

Takeaway for NSVD: According to the recent advances in NSVD, we have the following insights:

- *External Knowledge Enhances Results.* Utilizing external knowledge bases in NSVD processes significantly enhances new slot value identification. These resources provide a rich contextual backdrop that aids models in accurately recognizing and categorizing new slot values, even in complex or ambiguous contexts.

- *Effective Knowledge Transfer Influences NSVD.* Implementing effective knowledge transfer mechanisms that connect known slots and values with new slots and values enhances the ability of NSVD models. It leverages existing slot knowledge to inform and guide the identification and integration of new slots and values, reducing the learning curve and improving the system's adaptability to dynamic conversational contexts.

5 Conclusion and Future Directions

This paper presents the first comprehensive survey of recent advances in OnExp. We begin by formulating the task, detailing representative data resources and evaluation protocols used. We then examine prevalent OnExp methods, including NID, NSVD, and Joint OnExp. Despite significant progress achieved, several challenges remain, inspiring promising frontiers for future research.

Early OnExp. Existing studies primarily concentrate on developing models to expand predefined ontologies using extensive utterances. Yet, real-world conversational agents necessitate the ability to rapidly recognize and adapt to evolving user needs and dialogue contexts (Li et al., 2023a,b),

thus highlighting the critical importance of early-stage OnExp. Early OnExp faces the unique challenge of identifying new ontological items with minimal utterances when a known ontology has been established using extensive data. In such a scenario, nascent ontological items risk being submerged by more prevalent ones. Although Liang and Liao (2023) showcased the effectiveness of CsePL in early intent discovery, more specific methods that fully address the unique challenges of this area remain largely under-explored. This highlights its significant potential as a promising field for future research.

Multi-modal OnExp. Current OnExp tasks generally learned new ontological items from purely text-modal utterances. However, practical interactions with conversational agents typically occur in multi-modal settings (Liao et al., 2018; Zhang et al., 2019; Wu et al., 2022b), suggesting that such multi-modal data can enhance new ontology learning. For example, incorporating visual data in e-commerce or audio cues in customer support could provide deeper contextual insights than text-only systems (Zhu et al., 2020). Despite its potential, multi-modal OnExp is still in its early stages, with limited research on effectively synergizing different modalities to expand ontologies. This emerging area promises to significantly improve the capabilities of conversational agents across different applications, necessitating more comprehensive research into advanced modality integration techniques and benchmarks of multi-modal data in OnExp.

Holistic OnExp. Prior OnExp research has mainly confined their ontology analyses to the CU module of conversational agents, assessing their performance via metrics such as recognition accuracy. This narrow focus, however, overlooks the broader impact of OnExp results on the other pivotal components of conversational agents, *e.g.*, dialogue management and response generation. Additionally, the rationality of newly expanded ontologies has seldom been thoroughly examined, raising questions about whether OnExp outcomes can genuinely enhance dialogue policy learning or the quality of generated responses. To fill these gaps, there is a compelling need for more integrated approaches in OnExp. These methods should extend beyond merely identifying new ontological items, to a thorough evaluation of their holistic impact on the entire conversational agents, ensuring that advancements in OnExp positively contribute to the evolution of conversational AI and improve both system performance and user interaction quality.

Limitations

This survey provides a comprehensive overview of the latest studies in OnExp. Despite our diligent efforts, some limitations may still persist:

Categorization. The survey makes the first attempt to organize the recent OnExp works into three distinct dimensions. This organization reflects our subjective interpretation and understanding. External insights on this categorization might enrich the perspectives presented.

Descriptions. The descriptions of the introduced OnExp approaches in this survey are kept highly succinct to allow broad coverage within the constraints of page limits. We intend for this survey to act as a starting point, directing readers to the original works for more detailed information.

Experimental Results. The leaderboard in this survey predominantly emphasizes broad comparisons of different OnExp approaches, such as the overarching system performance, instead of detailed analyses. Going forward, we aim to expand on these comparisons with more in-depth analyses of the experimental outcomes, thereby offering a more comprehensive understanding of the strengths and weaknesses of various OnExp models.

Acknowledgments

References

Luca Maria Aiello, Debora Donato, Umut Ozertem, and Filippo Menczer. 2011. Behavior-driven clustering of queries into topics. In *CIKM*, pages 1373–1382.

Wenbin An, Wenkai Shi, Feng Tian, Haonan Lin, Qianying Wang, Yaqiang Wu, Mingxiang Cai, Luyan Wang, Yan Chen, Haiping Zhu, and Ping Chen. 2023a. Generalized category discovery with large language models in the loop. *CoRR*.

Wenbin An, Feng Tian, Ping Chen, Siliang Tang, Qinghua Zheng, and Qianying Wang. 2022. Fine-grained category discovery under coarse-grained supervision with hierarchical weighted self-contrastive learning. In *EMNLP*, pages 1314–1323.

Wenbin An, Feng Tian, Wenkai Shi, Yan Chen, Yaqiang Wu, Qianying Wang, and Ping Chen. 2024. Transfer and alignment network for generalized category discovery. In *AAAI*, pages 10856–10864.

Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, Qianying Wang, and Ping Chen. 2023b. Generalized category discovery with decoupled prototypical network. In *AAAI*, pages 12527–12535.

Avishek Anand, V. Venktesh, Abhijit Anand, and Vinay Setty. 2023. Query understanding in the age of large language models. *ArXiv*.

Ankur Bapna, Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2017. Towards zero-shot frame semantic parsing for domain scaling. In *INTERSPEECH*, pages 2476–2480.

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. 2004. Active semi-supervision for pairwise constrained clustering. In *ICDM*, pages 333–344.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *EMNLP*, pages 5016–5026.

David Carmel, Liane Lewin-Eytan, and Yoelle Maarek. 2018. Product question answering using customer generated content - research challenges. In *SIGIR*, pages 1349–1350.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 139–156.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *NLP4ConvAI@ACL*, pages 38–45.

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2017. Deep adaptive image clustering. In *ICCV*, pages 5880–5888.

Lingzhen Chen and Alessandro Moschitti. 2019. Transfer learning for sequence labeling using source model and target data. In *AAAI*, pages 6260–6267.

Yun-Nung Chen, William Yang Wang, and Alexander Rudnicky. 2015. Jointly modeling inter-slot relations by random walk on knowledge graphs for unsupervised spoken language understanding. In *NAACL*, pages 619–629.

Yun-Nung Chen, William Yang Wang, and Alexander I. Rudnicky. 2013. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 120–125.

Yun-Nung Chen, William Yang Wang, and Alexander I Rudnicky. 2014. Leveraging frame semantics and distributional semantics for unsupervised semantic slot induction in spoken dialogue systems. In *SLT*, pages 584–589.

Jackie Chi Kit Cheung and Xiao Li. 2012. Sequence clustering and labeling for unsupervised query intent discovery. In *WSDM*, pages 383–392.

Sam Coope, Tyler Farghly, Daniela Gerz, Ivan Vulic, and Matthew Henderson. 2020. Span-convert: Few-shot span extraction for dialog with pretrained conversational representations. In *ACL*, pages 107–121.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*.

Huy Dao, Yang Deng, Dung D. Le, and Lizi Liao. 2024. Broadening the view: Demonstration-augmented prompt learning for conversational recommendation. In *SIGIR*, pages 785–795.

Huy Dao, Lizi Liao, Dung Le, and Yuxiang Nie. 2023. Reinforced target-driven conversational promotion. In *EMNLP*, pages 12583–12596.

Maarten De Raedt, Fréderic Godin, Thomas Demeester, and Chris Develder. 2023. IDAS: Intent discovery with abstractive summarization. In *NLP4ConvAI@ACL*, pages 71–88.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Xinya Du, Luheng He, Qi Li, Dian Yu, Panupong Pasupat, and Yuan Zhang. 2021. Qa-driven zero-shot slot filling with weak supervision pretraining. In *ACL/IJCNLP*, pages 654–664.

Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *ACL*, pages 5467–5471.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Kumar Goyal, Peter Ku, and Dilek Hakkani-Tür. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *LREC*, pages 422–428.

K. Chidananda Gowda and G. Krishna. 1978. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognit.*, pages 105–112.

Shasha Guo, Lizi Liao, Jing Zhang, Yanling Wang, Cuiping Li, and Hong Chen. 2024. SGSH: stimulate large language models with skeleton heuristics for knowledge base question generation. In *Findings of NAACL*, pages 4613–4625.

Dilek Hakkani-Tür, Asli Celikyilmaz, Larry P. Heck, and Gökhan Tür. 2013. A weakly-supervised approach for discovering new user intents from search query logs. In *INTERSPEECH*, pages 3780–3784.

Dilek Hakkani-Tür, Yun-Cheng Ju, Geoffrey Zweig, and Gökhan Tür. 2015. Clustering novel intents in a conversational interaction system with semantic parsing. In *INTERSPEECH*, pages 1854–1858.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2019. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, pages 8400–8408.

Keqing He, Yuanmeng Yan, Hong Xu, Sihong Liu, Zijun Liu, and Weiran Xu. 2020a. Learning label-relational output structure for adaptive sequence labeling. In *IJCNN*, pages 1–8.

Keqing He, Yuanmeng Yan, and Weiran Xu. 2020b. Learning to tag OOV tokens by integrating contextual representation and background knowledge. In *ACL*, pages 619–624.

Keqing He, Jinchao Zhang, Yuanmeng Yan, Weiran Xu, Cheng Niu, and Jie Zhou. 2020c. Contrastive zero-shot learning for cross-domain slot filling with adversarial attack. In *ACL*, pages 1461–1467.

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

Matthew Henderson, Milica Gasic, Blaise Thomson, Pirros Tsiakoulis, Kai Yu, and Steve J. Young. 2012. Discriminative spoken language understanding using word confusion networks. In *SLT*, pages 176–181.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *ACL*, pages 1381–1393.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. 2018. Learning to cluster in order to transfer across domains and tasks. In *ICLR*.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. 2019. Multi-class classification without multi-class labels. In *ICLR*.

Ziniu Hu, Ting Chen, Kai-Wei Chang, and Yizhou Sun. 2019. Few-shot representation learning for out-of-vocabulary words. In *ACL*, pages 4102–4112.

Vojtěch Hudeček, Ondřej Dušek, and Zhou Yu. 2021. Discovering dialogue slots with weak supervision. In *ACL-IJCNLP*, pages 2430–2442.

Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manag.*, pages 1251–1266.

Rajat Kumar, Mayur Patidar, Vaibhav Varshney, Lovekesh Vig, and Gautam Shroff. 2022. Intent detection and discovery from user logs via deep semi-supervised contrastive clustering. In *NAACL-HLT*, pages 1836–1853.

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

Stefan Larson and Kevin Leach. 2022. A survey of intent classification and slot-filling datasets for task-oriented dialog. *CoRR*, abs/2207.13211.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *EMNLP-IJCNLP*, pages 1311–1316.

Sungjin Lee and Rahul Jha. 2019. Zero-shot adaptive transfer for conversational language understanding. In *AAAI*, pages 6642–6649.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2023a. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems (TOIS)*.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, et al. 2022. Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. In *ACL*.

Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *AAAI*, pages 18462–18470.

Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. 2023b. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *ACM MM*, pages 5923–5934.

Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, Guwei Jin, and Wei Chu. 2017. *AliMe Assist* : An intelligent assistant for creating an innovative e-commerce experience. In *CIKM*, pages 2495–2498.

Xuefeng Li, Liwen Wang, Guanting Dong, Keqing He, Jinzheng Zhao, Hao Lei, Jiachi Liu, and Weiran Xu. 2023c. Generative zero-shot prompt learning for cross-domain slot filling with inverse prompting. In *Findings of ACL*, pages 825–834.

Dongyun Liang, Weiran Xu, and Yinge Zhao. 2017. Combining word-level and character-level representations for relation classification of informal text. In *Rep4NLP@ACL*, pages 43–47.

Jinggui Liang and Lizi Liao. 2023. Clusterprompt: Cluster semantic enhanced prompt learning for new intent discovery. In *Findings of EMNLP*, pages 10468–10481.

Jinggui Liang, Lizi Liao, Hao Fei, and Jing Jiang. 2024a. Synergizing large language models and pre-trained smaller models for conversational intent discovery. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14133–14147.

Jinggui Liang, Lizi Liao, Hao Fei, Bobo Li, and Jing Jiang. 2024b. Actively learn from llms with uncertainty propagation for generalized category discovery. In *NAACL-HLT*.

Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 801–809.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *AAAI*, pages 8360–8367.

Han Liu, Xiaotong Zhang, Lu Fan, Xuandi Fu, Qimai Li, Xiao-Ming Wu, and Albert Y. S. Lam. 2019. Reconstructing capsule networks for zero-shot intent classification. In *EMNLP-IJCNLP*, pages 4798–4808.

Han Liu, Siyang Zhao, Xiaotong Zhang, Feng Zhang, Junjie Sun, Hong Yu, and Xianchao Zhang. 2022a. A simple meta-learning paradigm for zero-shot intent classification with mixture attention mechanism. In *SIGIR*, pages 2047–2052.

Jian Liu, Mengshi Yu, Yufeng Chen, and Jinan Xu. 2022b. Cross-domain slot filling as machine reading comprehension: A new perspective. *IEEE ACM Trans. Audio Speech Lang. Process.*, pages 673–685.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Dragomir R. Radev, and Arman Cohan. 2023. On learning to summarize with large language models as references. *ArXiv*.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020. Coach: A coarse-to-fine approach for cross-domain slot filling. In *ACL*, pages 19–25.

Qiaoyang Luo and Lingqiao Liu. 2023. Zero-shot slot filling with slot-prefix prompting and attention relationship descriptor. In *AAAI*.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297.

Yutao Mou, Keqing He, Pei Wang, Yanan Wu, Jingang Wang, Wei Wu, and Weiran Xu. 2022a. Watch the neighbors: A unified k-nearest neighbor contrastive learning framework for OOD intent discovery. In *EMNLP*, pages 1517–1529.

Yutao Mou, Keqing He, Yanan Wu, Zhiyuan Zeng, Hong Xu, Huixing Jiang, Wei Wu, and Weiran Xu. 2022b. Disentangled knowledge transfer for OOD intent discovery with unified contrastive learning. In *ACL*, pages 46–53.

Yutao Mou, Xiaoshuai Song, Keqing He, Chen Zeng, Pei Wang, Jingang Wang, Yunsen Xian, and Weiran Xu. 2023. Decoupling pseudo label disambiguation and representation learning for generalized intent discovery. In *ACL*, pages 9661–9675.

Nikola Mrksic, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *ACL*, pages 1777–1788.

Danilo Neves Ribeiro, Jack Goetz, Omid Abdar, Mike Ross, Annie Dong, Kenneth Forbus, and Ahmed Mohamed. 2023. Towards zero-shot frame semantic parsing with task agnostic ontologies and simple labels. In *Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, pages 54–63.

Hoang Nguyen, Chenwei Zhang, Ye Liu, and Philip Yu. 2023. Slot induction via pre-trained language model probing and multi-level contrastive learning. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 470–481, Prague, Czechia. Association for Computational Linguistics.

Cennet Oguz and Ngoc Thang Vu. 2021. Few-shot learning for slot tagging with attentive relational network. In *EACL*, pages 1566–1572.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Soham Parikh, Mitul Tiwari, Prashil Tumbade, and Quaizar Vohra. 2023. Exploring zero and few-shot techniques for intent classification. In *ACL*, pages 744–751.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI*, pages 8689–8696.

Xiang Ren, Yujing Wang, Xiao Yu, Jun Yan, Zheng Chen, and Jiawei Han. 2014. Heterogeneous graph-based intent learning with queries, web pages and wikipedia concepts. In *WSDM*, pages 23–32.

Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in web search. In *WWW*, pages 13–19.

Darsh J. Shah, Raghav Gupta, Amir A. Fayazi, and Dilek Hakkani-Tür. 2019. Robust zero-shot cross-domain slot filling with example values. In *ACL*, pages 5484–5490.

Xiang Shen, Yinge Sun, Yao Zhang, and Mani Najmabadi. 2021. Semi-supervised intent discovery with contrastive learning. In *NLP4CONVAI*, pages 120–129.

Wenkai Shi, Wenbin An, Feng Tian, Qinghua Zheng, Qianying Wang, and Ping Chen. 2023. A diffusion weighted graph framework for new intent discovery. In *EMNLP*, pages 8033–8042.

Qingyi Si, Yuanxin Liu, Peng Fu, Zheng Lin, Jiangnan Li, and Weiping Wang. 2021. Learning class-transductive intent representations for zero-shot intent detection. In *IJCAI*, pages 3922–3928.

A. B. Siddique, Fuad T. Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized zero-shot intent detection via commonsense knowledge. In *SIGIR*, pages 1925–1929.

Xiaoshuai Song, Keqing He, Pei Wang, Guanting Dong, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2023. Large language models meet open-world intent discovery and recognition: An evaluation of chatgpt. In *EMNLP*, pages 10291–10304.

Mujeen Sung, James Gung, Elman Mansimov, Nikolaos Pappas, Raphael Shu, Salvatore Romeo, Yi Zhang, and Vittorio Castelli. 2023. Pre-training intent-aware encoders for zero- and few-shot intent classification. In *Proceedings of EMNLP*, Singapore.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Gökhan Tür, Dilek Hakkani-Tür, Dustin Hillard, and Asli Celikyilmaz. 2011. Towards unsupervised spoken language understanding: Exploiting query click logs for slot filling. In *INTERSPEECH*, pages 1293–1296.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2022. Generalized category discovery. In *CVPR*, pages 7482–7491.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large language models enable few-shot clustering.

Jixuan Wang, Kai Wei, Martin Radfar, Weiwei Zhang, and Clement Chung. 2021a. Encoding syntactic knowledge in transformer encoder for intent detection and slot filling. In *AAAI*, pages 13943–13951.

Liwen Wang, Xuefeng Li, Jiachi Liu, Keqing He, Yuanmeng Yan, and Weiran Xu. 2021b. Bridge to target domain by prototypical contrastive learning and label confusion: Re-explore zero-shot learning for slot filling. In *EMNLP*, pages 9474–9480.

Pei Wang, Keqing He, Yejie Wang, Xiaoshuai Song, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2024. Beyond the known: Investigating llms performance on out-of-domain intent detection. *CoRR*.

Xueying Wang, Haiqiao Zhang, Qi Li, Yiyu Shi, and Meng Jiang. 2019. A novel unsupervised approach for precise temporal slot filling from incomplete and noisy temporal contexts. In *WWW*, pages 3328–3334.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve J. Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *EACL*, pages 438–449.

Ting-Wei Wu, Ruolin Su, and Biing-Hwang Juang. 2021. A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding. In *EMNLP*, pages 4884–4896.

Yuxia Wu, Tianhao Dai, Zhedong Zheng, and Lizi Liao. 2024. Active discovering new slots for task-oriented conversation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yuxia Wu, Lizi Liao, Xueming Qian, and Tat-Seng Chua. 2022a. Semi-supervised new slot discovery with incremental clustering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6207–6218.

Yuxia Wu, Lizi Liao, Gangyi Zhang, Wenqiang Lei, Guoshuai Zhao, Xueming Qian, and Tat-Seng Chua. 2022b. State graph reasoning for multimodal conversational recommendation. *IEEE Transactions on Multimedia*, 25:3113–3124.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *EMNLP*, pages 3090–3099.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*, pages 478–487.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *VS@HLT-NAACL*, pages 62–69.

Zengzhuang Xu, Rui Song, Bowei Zou, and Yu Hong. 2017. Unsupervised slot filler refinement via entity community construction. In *NLPCC*, pages 642–651.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y. S. Lam. 2020. Unknown intent detection using gaussian mixture model with an application to zero-shot intent classification. In *ACL*, pages 1050–1060.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *ICML*, pages 3861–3870.

Dian Yu and Heng Ji. 2016. Unsupervised person slot filling based on graph mining. In *ACL*, pages 44–53.

Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent Shafey, and Hagen Soltau. 2022. Unsupervised slot schema induction for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1193.

Zengfeng Zeng, Dan Ma, Haiqin Yang, Zhen Gou, and Jianping Shen. 2021. Automatic intent-slot induction for dialogue systems. In *WWW*, pages 2578–2589.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021a. Supporting clustering with contrastive learning. In *NAACL-HLT*, pages 5419–5430.

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. 2021b. TEXTOIR: An integrated and visualized platform for text open intent recognition. In *ACL (demo)*, pages 167–174.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021c. Discovering new intents with deep aligned clustering. In *AAAI*, pages 14365–14373.

Hanlei Zhang, Huanlin Xu, Xin Wang, Fei Long, and Kai Gao. 2023a. A clustering framework for unsupervised and semi-supervised new intent discovery. *TKDE*.

Junwen Zhang and Yin Zhang. 2023. Hierarchicalcontrast: A coarse-to-fine contrastive learning framework for cross-domain zero-shot slot filling. In *Findings of EMNLP*, pages 14483–14503.

Shun Zhang, Jian Yang, Jiaqi Bai, Chaoran Yan, Tongliang Li, Zhao Yan, and Zhoujun Li. 2024. New intent discovery with attracting and dispersing prototype. In *LREC-COLING*, pages 12193–12206.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023b. Clusterllm: Large language models as a guide for text clustering. In *EMNLP*, pages 13903–13920.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Y. S. Lam. 2022. New intent discovery with pre-training and contrastive learning. In *ACL*, pages 256–269.

Zheng Zhang, Lizi Liao, Minlie Huang, Xiaoyan Zhu, and Tat-Seng Chua. 2019. Neural multimodal belief tracker with adaptive attention for dialogue systems. In *The world wide web conference*, pages 2401–2412.

Lin Zhao and Zhe Feng. 2018. Improving slot filling in spoken language understanding with joint pointer and attention. In *ACL*, pages 426–431.

Yunhua Zhou, Guofeng Quan, and Xipeng Qiu. 2023. A probabilistic framework for discovering new intents. In *ACL*, pages 3771–3784.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal joint attribute prediction and value extraction for e-commerce product. In *EMNLP*, pages 2129–2139.

A Appendix

A.1 Data Resources

New Intent Discovery Datasets. We show three widely used datasets for NID. Specifically, **BANG-ING77** (Casanueva et al., 2020) is a fine-grained intent discovery dataset sourced from banking domain dialogues. It contains over 13K user utterances distributed across 77 unique intents. **CLINC150** (Larson et al., 2019), on the other hand, is a multi-domain dataset featuring 150 distinct intents and 22,500 utterances across 10 different domains. **StackOverflow** (Xu et al., 2015), a dataset curated from Kaggle.com, includes 20,000 technical questions categorized into 20 distinct areas.

New Slot-Value Discovery Datasets. For the NSVD task, we introduce seven prominent datasets spanning various domains. The **CamRest** dataset, provided by Wen et al. (2017), delves into the restaurant domain, boasting over 2,700 utterances across 4 slots, offering valuable insights into task-oriented dialogues. Similarly, the **Cambridge SLU** dataset by Henderson et al. (2012) also explores the restaurant sector, featuring more than 10,500 utterances across 5 slots. Additionally, the MultiWOZ dataset spans multiple domains, with its subsets, **WOZ-attr** (Eric et al., 2020) and **WOZ-hotel** (Eric et al., 2020), exploring the attraction and hotel domains with over 7,500 and 14,000 utterances, respectively. Despite encompassing intents, the limited intent quantity in these datasets restricts their suitability for the NID task. Conversely, the **ATIS** dataset (Hemphill et al., 1990) expands into the flight domain with nearly 5,000 utterances and 120 slots. The **SNIPS** dataset (Coucke et al., 2018) provides a valuable resource for spoken language understanding across seven domains, boasting 72 slots and around 2,000 utterances per domain. The **SGD** (Rastogi et al., 2020) contains dialogues from 16 domains with a total of 46 intents and 214 slots. Notably, ATIS, SNIPS, and SGD are replete with a variety of intents, thus making them apt for comprehensive studies in both NID and NSVD tasks.

A.2 Evaluation Protocols

NID Metrics. The NID task involves accurately assigning utterances into their corresponding intent groups from potentially many possibilities. Accordingly, the performance of NID models is typically assessed using three standard metrics: ACC, ARI, and NMI (Zhang et al., 2021c, 2022), which evalu-

ate how effectively the model identifies and groups intents, ensuring that the clustering reflects true user intentions rather than random associations. As previously mentioned, ACC assesses NID performance by calculating the proportion of correctly predicted outputs to total predictions, aligned with ground-truth labels. Notably, the ACC in this context is derived following an alignment process using the Hungarian algorithm. The definition of ACC is as follows:

$$ACC = \frac{\sum_{i=1}^{N} \mathbb{1}_{y_i = map(\hat{y}_i)}}{N}, \qquad (2)$$

where $\{\hat{y}_i, y_i\}$ denote the predicted and true labels, respectively. $map(\cdot)$ is the Hungarian algorithm-based mapping function.

Different from ACC, ARI measures the concordance of the predicted and actual clusters through an assessment of pairwise accuracy within clusters, which is computed as:

$$ARI = \frac{\sum_{i,j} \binom{n_{i,j}}{2} - [\sum_i \binom{u_i}{2} \sum_j \binom{v_j}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_i \binom{u_i}{2} + \sum_j \binom{v_j}{2}] - [\sum_i \binom{u_i}{2} \sum_j \binom{v_j}{2}]/\binom{N}{2}}, \qquad (3)$$

where $n_{i,j}$ denotes the number of sample pairs both in i^{th} predicted and j^{th} ground-truth cluster. $u_i = \sum_j n_{i,j}$, and $v_j = \sum_i n_{i,j}$ represent the sum of sample pairs in the same predicted and true clusters, respectively. N is the number of all samples.

Regarding the NMI, it aims to gauge the level of agreement between the predicted and ground-truth clusters by quantifying the normalized mutual information between them. It can be calculated as follows:

$$NMI(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{2 \cdot I(\hat{\boldsymbol{y}}, \boldsymbol{y})}{H(\hat{\boldsymbol{y}}) + H(\boldsymbol{y})}, \qquad (4)$$

where $\{\hat{\boldsymbol{y}}, \boldsymbol{y}\}$ denote the predicted labels and the ground-truth labels respectively. $I(\cdot)$ signifies mutual information. $H(\cdot)$ is the entropy function.

NSVD Metrics. For the NSVD task, the challenge lies in accurately identifying relevant slots and values within utterances and precisely delineating their boundaries. Metrics such as Precision, Recall, and Span-F1 are essential for assessing the performance of NSVD models. These metrics ensure the accuracy and completeness of information extraction, focusing on specific elements within utterances. Considering a set of actual slot values $M_1, M_2, \ldots, M_n$, where n is the number of slots, and a corresponding set of predicted values

Methods	BANKING77			CLINC150			StackOverflow		
	ACC	ARI	NMI	**ACC**	ARI	NMI	**ACC**	ARI	NMI
Statistical Methods									
K-Means (MacQueen et al., 1967)	29.55	12.18	54.57	45.06	26.86	70.89	13.55	1.46	8.24
AG (Gowda and Krishna, 1978)	31.58	13.31	57.07	44.03	27.70	73.07	14.66	2.12	10.62
NN-based Methods									
DEC (Xie et al., 2016)	41.29	27.21	67.78	46.89	27.46	74.83	13.09	3.76	10.88
DCN (Yang et al., 2017)	41.99	26.81	67.54	49.29	31.15	75.66	34.26	15.45	31.09
DAC (Chang et al., 2017)	27.41	14.24	47.35	55.94	40.49	78.40	16.30	2.76	14.71
DeepCluster (Caron et al., 2018)	20.69	8.95	41.77	35.70	19.11	65.58	-	-	-
SCCL (Zhang et al., 2021a)	40.54	26.98	63.89	50.44	38.14	79.35	68.15	34.81	69.11
USNID (Zhang et al., 2023a)	54.83	43.33	75.30	75.87	68.54	91.00	69.28	52.25	72.00
IDAS (De Raedt et al., 2023)	67.43	57.56	82.84	85.48	79.02	93.82	83.82	72.20	81.26

Table 4: The main unsupervised NID results on three benchmarks.

Methods	CamRest	Cambridge SLU	WOZ-hotel	WOZ-attr	ATIS
DistFrame-Sem(Chen et al., 2014)	53.5	59.0	38.2	37.5	61.6
Merge-Select(Hudeček et al., 2021)	55.2	66.4	38.8	38.3	64.8

Table 5: The main results of unsupervised NSVD methods on five benchmarks. Here we provide the Span-F1 score.

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, precision P_i and recall R_i are calculated for each slot type i as follows:

$$P_i = \frac{|M_i \cap \varepsilon_i|}{|\varepsilon_i|}, \tag{5}$$

$$R_i = \frac{|M_i \cap \varepsilon_i|}{|M_i|}. \tag{6}$$

The overall weighted precision P and recall R are computed as follows:

$$P = \frac{\sum_{i=1}^n |\varepsilon_i| P_i}{\sum_{j=1}^n |\varepsilon_j|}, \tag{7}$$

$$R = \frac{\sum_{i=1}^n |M_i| R_i}{\sum_{j=1}^n |M_j|}. \tag{8}$$

The F1 score is then computed as the harmonic mean of the overall weighted precision and recall, thus accounting for both the precision and recall in a balanced manner:

$$F1 = \frac{2PR}{P + R}. \tag{9}$$

In the context of slot value spans, this metric is specifically referred to as Span-F1.

Other Metrics. While NID and NSVD metrics offer valuable insights into OnExp model performance, their uniform application across all test data can obscure distinctions between utterances containing known versus novel ontological items. To address this, metrics such as **Known ACC**, **Novel ACC**, and the **H-score** are indispensable, as they effectively differentiate model performance on known and novel items, providing a more granular assessment of model capabilities (An et al., 2024). Specifically, **Known ACC** and **Novel ACC** are specialized forms of **ACC**, computed separately for known and novel ontological items. The **H-score** is calculated as the harmonic mean of **Known ACC** and **Novel ACC** as follows:

$$\text{H-score} = \frac{2}{1/\text{Know ACC} + 1/\text{Novel ACC}}. \tag{10}$$

B Leaderboard

NID Leaderboard. Table 4 presents the unsupervised NID results on three benchmarks. Notably, although USNID is categorized into the semi-supervised NID methods, it can adapt to an unsupervised setting. Hence, we have included USNID results in the unsupervised context for a comprehensive evaluation.

NSVD Leaderboard. Table 5 and Table 6 present the main performance of unsupervised NSVD

Domain	Sequence tagging-based models					MRC-based models	Prompting-based models
	CT	RZT	Coach	CZSL	PCLC	RCSF	GZPL
AddToPlaylist	38.82	42.77	50.90	53.89	59.24	68.70	61.64
BookRestaurant	27.54	30.68	34.01	34.06	41.36	63.49	62.93
GetWeather	46.45	50.28	50.47	52.04	54.21	65.36	64.97
PlayMusic	32.86	33.12	32.01	34.59	34.95	53.51	66.42
RateBook	14.54	16.43	22.06	31.53	29.31	36.51	47.53
SearchCreativeWork	39.79	44.45	46.65	50.61	53.51	69.22	72.88
SearchScreeningEvent	13.83	12.25	25.63	30.05	27.17	33.54	51.42
Average F1	30.55	32.85	37.39	40.99	42.82	55.76	61.07

Table 6: The main results of Partially Supervised NSVD methods the SNIPS dataset.

methods and partially supervised NSVD methods.

We adopted results reported in the published literature (Zhang et al., 2021c, 2023a; Zhou et al., 2023; Zhang et al., 2024; Liang and Liao, 2023; Wu et al., 2022a, 2024).