HiCoLoRA: Addressing Context-Prompt Misalignment via Hierarchical Collaborative LoRA for Zero-Shot DST

Anonymous ACL submission

Abstract

Zero-shot Dialog State Tracking (zs-DST) is essential for enabling Task-Oriented Dialog Systems (TODs) to generalize to new domains without costly data annotation. A central challenge lies in the semantic misalignment between dynamic dialog contexts and static prompts, leading to inflexible cross-layer coordination, domain interference, and catastrophic forgetting. To tackle this, we propose Hierarchical Collaborative Low-Rank Adaptation (HiCoLoRA), a framework that enhances zero-shot slot inference through robust prompt alignment. It features a hierarchical LoRA architecture for dynamic layer-specific processing (combining lower-layer heuristic grouping and higher-layer full interaction), integrates Spectral Joint Domain-Slot Clustering to identify transferable associations (feeding an Adaptive Linear Fusion Mechanism), and employs Semantic-Enhanced SVD Initialization (SemSVD-Init) to preserve pre-trained knowledge. Experiments on multi-domain datasets MultiWOZ and SGD show that HiCoLoRA outperforms baselines, achieving SOTA in zs-DST. Code is available at Anonymous Github.

1 Introduction

Task-Oriented Dialog Systems (TODs) assist users to complete specific tasks, such as hotel or flight bookings (Luo et al., 2024; Wang et al., 2024d). To generalize to new domains without extensive training, TODs rely on zero-shot Dialog State Tracking (zs-DST), yet face challenges of data scarcity and domain adaptation. A central obstacle is the semantic misalignment between dynamic dialog contexts and static prompts. This misalignment limits effective cross domain transfer.

To address data scarcity and enable generalization, existing approaches for zs-DST include data augmentation (He et al., 2025), prompt engineering (Liu et al., 2025b; Wang et al., 2024c; Aksu et al., 2023), and parameter-efficient fine-tuning (PEFT). Among PEFT methods, Low-Rank Adaptation (LoRA) (Wang et al., 2024a; Occhipinti et al., 2024) has gained prominence for zs-DST (Yi et al., 2025; Aksu et al., 2023) due to its ability to adapt large models with minimal parameters. Building upon LoRA, recent multi-LoRA variants such as DualLoRA (Luo et al., 2024), CoLA (Zhou et al., 2025), HydraLoRA (Tian et al., 2024), and MTL-LoRA (Yang et al., 2025) enhance adaptability through specialized adapters or cross-task collaboration. The semantic misalignment arises from the inherent layer-wise processing in Transformers: lower layers capture local semantic atoms while higher layers integrate them into global intent representations (Liu et al., 2024). Despite these advances, existing LoRA variants treat all layers uniformly, failing to coordinate these distinct semantic roles, especially in zero-shot scenarios where domain-specific cues are scarce.

These limitations stem from a structural mismatch between dynamic dialog contexts and static prompts as illustrated in Fig. 1, which manifests in three critical research questions: (RQ1) Rigid hierarchical designs hinder effective cross-layer weight sharing, limiting fine-grained semantic alignment in deeper layers. (RQ2) A single adaptation matrix conflates domain-agnostic and domain specific signals, causing semantic confusion between domains. (RQ3) The use of random initialization for LoRA parameters can distort pre-trained knowledge and exacerbate catastrophic forgetting.

To address the three limitations, we propose Hierarchical Collaborative Low-Rank Adaptation (HiCoLoRA), a novel framework inspired by DualLoRA's prompt augmentation (Luo et al., 2024) and CoLA's multi-LoRA grouping (Zhou et al., 2025). Departing from "uniform layer processing", our contributions are: **(1)** A Hierarchical Collaborative Architecture with lower-layer heuristic grouping and higher-layer full interaction, resolving RQ1 via dynamic cross-layer coordination; **(2)** Spectral

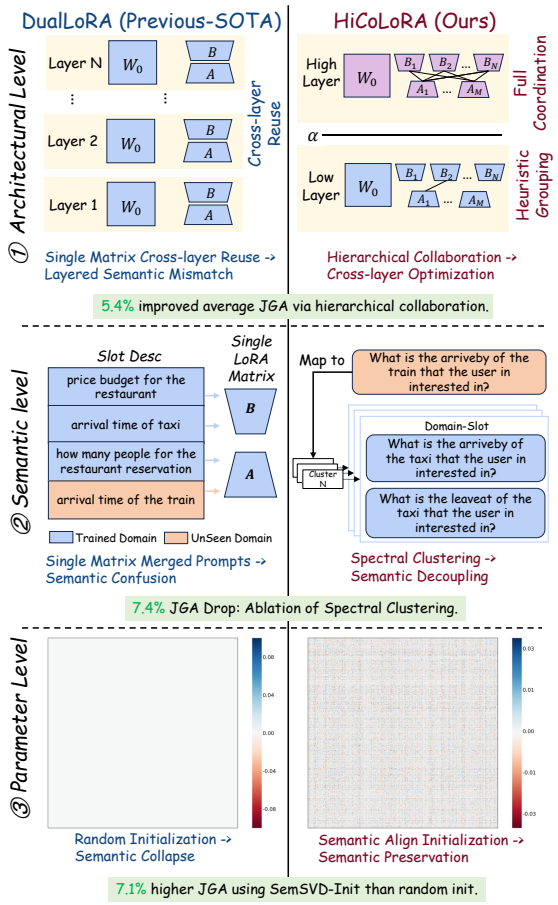


Figure 1: Three critical challenges motivating our work: (1) Architectural rigidity hinders cross-layer coordination in Transformers, limiting fine-grained semantic alignment; (2) Coupling of domain-shared and domain-specific semantics causes cross-domain confusion; (3) Random parameter initialization distorts pre-trained knowledge, exacerbating catastrophic forgetting.

Joint Clustering and Adaptive Fusion disentangling domain-shared and specific semantics addressing RQ2; **(3)** Semantic-Enhanced SVD Initialization preserving pre-trained knowledge against RQ3.

Extensive experiments on MultiWOZ and SGD multi-domain datasets demonstrate that HiCoLoRA significantly outperforms previous SOTA methods, validating the effectiveness of our proposed framework in addressing the core challenges of zs-DST.

2 Related Work

Layer-Specific Algorithms in Transformers. The hierarchical processing in Transformers is well established: lower layers capture local semantic atoms, while higher layers integrate them into global intent representations (Liu et al., 2024; Wang et al., 2025). To exploit this asymmetry, various layer specific algorithms have been proposed, including hierarchical LoRA (Xiao et al., 2024; Guo et al., 2024), dynamic layer replacement (Xiong et al., 2024), attention head pruning (He and Lin, 2025; Zayed et al., 2024), and split attention mechanisms (Lin et al., 2025). While these methods improve efficiency by leveraging unequal layer contributions, they typically treat layers uniformly, lacking mechanisms to coordinate their distinct semantic roles, particularly crucial for dynamic dialog contexts where fine grained alignment between evolving utterances and static prompts is required. HiCoLoRA addresses this gap through explicit hierarchical collaboration, enabling dynamic cross-layer coordination.

PEFT with LoRA. LoRA and its variants have become prominent for zs-DST due to their parameter efficiency (Zhang et al., 2025; Liu et al., 2025a; Jabbarvaziri and Lampe, 2025). DualLoRA (Luo et al., 2024) uses dual adapters to align contexts and prompts, while multi-adapter approaches like HydraLoRA (Tian et al., 2024), CoLA (Zhou et al., 2025), and MTL-LoRA (Yang et al., 2025) enhance cross-task collaboration. Methods like RoSA (Nikdan et al., 2024) combine low-rank and sparse adaptations for efficiency, while initialization strategies like PiSSA (Meng et al., 2024) and MiLoRA (Zhang et al., 2024) aim to better preserve pre-trained knowledge. Spectral based adaptations have also been explored (Zhang and Pilanci, 2025; Wang et al., 2024b). Yet, these approaches still struggle with the semantic misalignment between dynamic dialog contexts and static slot prompts, often conflating domain agnostic and domain specific signals, which hinders zero-shot generalization. HiCoLoRA directly tackles this via spectral joint clustering and adaptive fusion, enabling disentangled and aligned representations.

zs-DST and Goal Accuracy. zs-DST aims to generalize to unseen domains without annotated data. Early methods like TRADE (Wu et al., 2019) and SUMBT (Lee et al., 2019) relied on task-specific architectures. With the advent of PLMs, generation based approaches like SimpleTOD (Hosseini-Asl et al., 2020) and intent enhanced methods (Yi et al., 2025) have advanced the field. Prompt based methods like Prompter (Aksu et al., 2023) and dual adapter approaches like DualLoRA (Luo et al., 2024) further improve cross-domain transfer, while synthetic data methods like LUAS (Wang et al., 2024d) address data scarcity. However, these approaches remain limited by rigid layer processing,

insufficient semantic disentanglement, and knowledge distortion during adaptation. HiCoLoRA fundamentally optimizes these challenges through its hierarchical cross-layer coordination and spectral domain-slot disentanglement.

3 Method

We propose Hierarchical Collaborative Low-Rank Adaptation (HiCoLoRA, Fig. 2) to address the research questions: A hierarchical collaborative architecture resolves cross-layer rigidity (RQ1) by enabling dynamic layer specific coordination; Spectral joint domain-slot clustering that disentangles domain shared and domain specific semantics (RQ2), guiding adaptive fusion between domain-agnostic (UniRep-LoRA) and domain specific (SemAdapt-LoRA) representations; and Semantic Enhanced SVD Initialization (SemSVD-Init) that preserves pre-trained knowledge against catastrophic forgetting (RQ3). These components systematically mitigate context prompt misalignment and strengthen zero-shot generalization.

3.1 Universal Representation LoRA (UniRep-LoRA)

UniRep-LoRA is designed to efficiently capture domain-agnostic semantic information from the dialog context $\boldsymbol{x}_{ur}$, such as universal slots for time and location. By freezing the parameters of the pre-trained model $\boldsymbol{W}_0$ and updating only the low-rank matrices $\boldsymbol{B}_{ur}$ and $\boldsymbol{A}_{ur}$:

$$\boldsymbol{h}_{ur} = \boldsymbol{W}_0 x_{ur} + \boldsymbol{B}_{ur}\boldsymbol{A}_{ur}\boldsymbol{x}_{ur}. \quad (1)$$

UniRep-LoRA and SemAdapt-LoRA are combined via adaptive linear fusion, balancing general and domain-specific representations to mitigate context-prompt misalignment for zero-shot scenarios.

3.2 Semantic Adaptation LoRA (SemAdapt-LoRA)

Unlike UniRep-LoRA which captures universal features, SemAdapt-LoRA specializes in domain specific prompt optimization, dynamically adjusting their influence across domains (addressing RQ2). We first employ a Multi-Head Attention module to enhance semantic alignment between dialog contexts and slot descriptions: high frequency dialog words from the training set serve as $\boldsymbol{Q}$, while slot descriptions act as $\boldsymbol{K}$ and $\boldsymbol{V}$, yielding the aligned representation $\boldsymbol{x}_{sa}$. This allows different heads to capture diverse semantic correlations, providing richer local semantic inputs for subsequent hierarchical processing.

To enable fine grained adaptation, we introduce two sets of trainable low-rank matrices: $\boldsymbol{A}_{sa}^{m}|_{m=1}^{M}$ for domain common prompt encoding, and $\boldsymbol{B}_{sa}^{n}|_{n=1}^{N}$ for cluster specific domain-slot reconstruction. Here, M denotes the number of identified domain clusters, and N the number of semantic clusters for slot prompts. $\boldsymbol{A}_{sa}^{m}$ compresses high dimensional prompt semantics into a low-rank space, while $\boldsymbol{B}_{sa}^{n}$ reconstructs domain specific representations from these compressed features.

To ensure effective collaboration between $\boldsymbol{A}_{sa}^{m}$ and $\boldsymbol{B}_{sa}^{n}$ across layers (addressing RQ1), we design a cross-layer collaborative module that respects the distinct semantic roles of Transformer layers: lower layers capture *local semantic features* serving as semantic atoms, while higher layers model *global semantic features* and guide lower-layer feature extraction via attention based suppression of irrelevant associations. This hierarchical approach moves beyond uniform layer processing, forming a semantic chain from local cues to global intent.

Heuristic Grouping. For lower layers that encode local semantic atoms, we adopt heuristic grouping to efficiently aggregate semantically similar parameters, reducing irrelevant interactions and providing a precise foundation for higher layer processing:

$$\boldsymbol{h}_{sa} = \boldsymbol{W}_0^{l}\boldsymbol{x}_{sa} + N\boldsymbol{B}_{sa}^{*}M\boldsymbol{A}_{sa}^{*}\boldsymbol{x}_{sa}, \quad (2)$$

where $\boldsymbol{W}_0^{l}$ denotes lower-layer weights. Matrices $\boldsymbol{A}_{sa}^{*}$ and $\boldsymbol{B}_{sa}^{*}$ are selected based on cosine similarity between the input embeddings $\boldsymbol{x}_{sa}$ and the cluster centroids of domains ($\mathcal{D}^{M}$) and slot prompts ($\mathcal{X}^{N}$), respectively. During training, differentiable selection is performed via Gumbel-Softmax, while at inference, we use softmax for efficiency.

Full Collaboration. Higher layers integrate all local semantic atoms through full collaboration, capturing implicit cross slot associations, such as *train-arriveby* and *destination*. This process can suppress irrelevant signals via attention guidance from lower layers:

$$\boldsymbol{h}_{sa} = \boldsymbol{W}_0^{h}\boldsymbol{x}_{sa} + \sum_{n=1}^{N}\boldsymbol{B}_{sa}^{n}\sum_{m=1}^{M}\boldsymbol{A}_{sa}^{m}\boldsymbol{x}_{sa}, \quad (3)$$

where $\boldsymbol{W}_0^{h}$ denotes higher-layer weights.

3.3 Adaptive Linear Fusion Mechanism

To dynamically integratethe two LoRA modules, we introduce a gated fusion mechanism with a

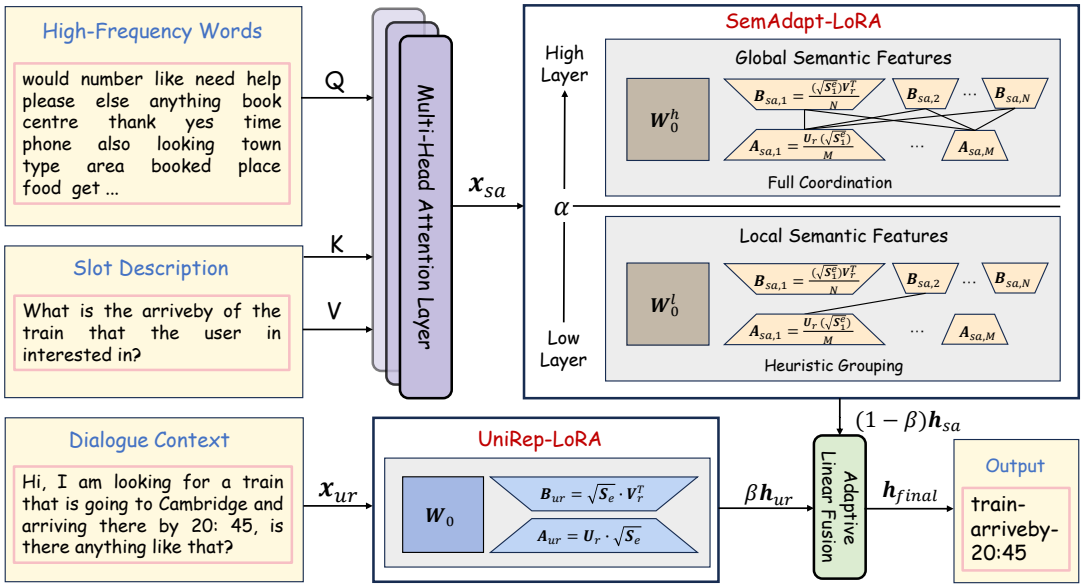


Figure 2: The HiCoLoRA framework combines: (1) UniRep-LoRA and SemAdapt-LoRA with Adaptive Linear Fusion balancing domain-agnostic and domain-specific features; (2) Spectral Joint Domain-Slot Clustering disentangling domain semantics to guide fusion; (3) SemSVD-Init preserving pre-trained knowledge via singular value modulation. These synergistically address context-prompt misalignment, enhancing zero-shot slot inference.

learnable coefficient β (trained end-to-end):

$$\boldsymbol{h}_{final} = \beta \boldsymbol{h}_{ur} + (1-\beta)\boldsymbol{h}_{sa}, \quad \beta \in (0,1). \quad (4)$$

Unlike static weighting, this adaptive gating allows the model to adjust the contribution of each module per turn based on the dialog context and slot descriptions, thereby mitigating the dynamic static prompt misalignment inherent in zs-DST.

3.4 Spectral Clustering of Domains and Slot Prompts

We propose a spectral joint clustering mechanism to capture semantic relatedness across domains and slot prompts, addressing the transferable associations needed for zero-shot adaptation. Domains often share abstract categories, such as *train* and *taxi* as transportation, while slot prompts are formatted as structured pairs *{domain-slot: question}*, for instance *{train-arriveby: what is the arrival time?}*. This reveals cross-domain semantic commonalities, such as *train-arriveby* and *taxi-arriveby* both expressing temporal attributes.

To cluster these semantically, we encode domain names and slot prompts using a T5 encoder into dense vector representations. Spectral clustering is then applied via Laplacian matrix eigendecomposition. The optimal cluster numbers (M for domains and N for slot prompts) are determined by maximizing the silhouette coefficient, resulting in clusters $\mathcal{D}^M$ and $\mathcal{X}^N$.

3.5 Semantic Enhanced SVD Initialization (SemSVD-Init)

SemSVD-Init initializes the low-rank matrices in both UniRep-LoRA and SemAdapt-LoRA by modulating the singular values of the pre-trained weight matrix according to domain-slot semantics. While Kaiming initialization may disrupt pre-trained representations; and PiSSA (Meng et al., 2024) lacks explicit task alignment, SemSVD-Init explicitly preserves pre-trained knowledge while enhancing semantics relevant to domains and slots, thereby directly addressing RQ3.

The core idea is to amplify singular directions corresponding to transferable semantic patterns and suppress those associated with domain specific noise, effectively aligning the parameter space with the clustered semantic structure.

Taking UniRep-LoRA as an example, we first perform SVD on the pre-trained weight $\boldsymbol{W}_0$:

$$\boldsymbol{W}_0 = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^T. \quad (5)$$

Subsequently, a correlation matrix $\boldsymbol{R}$ is computed by cosine similarity between the right singular vectors $\boldsymbol{V}_r$ and the cluster embeddings $\text{T5}_{\text{en}}(\mathcal{X}^N)$, where T5_{en} denotes the embeddings

of the encoder of the T5 model.

$$\boldsymbol{R} = \cos(\boldsymbol{V}_r, \mathrm{T5}_{en}(\mathcal{X}^N)). \tag{6}$$

Using these correlations, the singular values are enhanced on the basis of maximum category relevance for each vector.

$$\boldsymbol{S}_e = \mathrm{diag}\left(\sigma_1 \cdot \mathrm{ReLU}(1 + \lambda \boldsymbol{R}_1), \ldots, \sigma_r \cdot \mathrm{ReLU}(1 + \lambda \boldsymbol{R}_r)\right), \tag{7}$$

where $\boldsymbol{R}_k$ is the relevance score for the k-th singular vector, derived from the correlation between $\boldsymbol{V}_r[:, k]$ and the cluster embeddings, $\mathrm{ReLU}(x) = \max(0, x)$ to ensure positivity, and λ is a hyperparameter. The LoRA matrices are initialized as:

$$\begin{aligned} \boldsymbol{A}_{ur} &= \sqrt{\boldsymbol{S}_e}\boldsymbol{V}_r^T, \\ \boldsymbol{B}_{ur} &= \boldsymbol{U}_r\sqrt{\boldsymbol{S}_e}. \end{aligned} \tag{8}$$

Finally, the residual weight matrix $\boldsymbol{W}_{res}$ is adjusted to preserve key knowledge of the pre-trained model and avoiding distortion of its semantic structure.

$$\boldsymbol{W}_{res} = \boldsymbol{W}_0 - \boldsymbol{B}_{ur}\boldsymbol{A}_{ur}. \tag{9}$$

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate on two standard multi-domain TOD benchmarks: MultiWOZ 2.1 and Schema-Guided Dialog (SGD). Both are split with strict domain separation for zero-shot evaluation (details in Appendix A.1).

Baseline. To evaluate the generalizability of the proposed HiCoLoRA method, we conduct a comparison against representative baselines and SOTA approaches with details in Appendix A.2. Additionally, comparisons with recent advanced LoRA variants and larger scale LLMs are included to thoroughly assess scalability and generalization.

Metrics. We evaluate all models using Joint Goal Accuracy (JGA) and Average Goal Accuracy (AGA). JGA measures the rate of turns with all slots exactly matched, indicating system-level reliability. AGA calculates the ratio of correctly predicted to total slots, accounting for missed true slots and errors, reflecting fine-grained slot recall and local semantic alignment. The metrics' formulas and additional experimental details are provided in Appendices A.3 and A.4.

4.2 Main Results and Analysis

We summarize the main findings from Table 1 (MultiWOZ) and Appendix Table 4 (SGD) below:

Overall Performance Superiority. HiCoLoRA achieves new SOTA results on both MultiWOZ and SGD, with an average JGA of 40.8 on MultiWOZ and consistent gains across all SGD domains. This improvement stems from our hierarchical adaptation design, which overcomes three key limitations of prior work: rigid feature engineering in traditional methods, catastrophic forgetting in full fine-tuning, and shallow or uniform adaptation in recent SOTA models. Notably, HiCoLoRA attains an AGA of 93.8% on SGD Trains, demonstrating its ability to preserve rare slot semantics through SemSVD-Init and maintain layer wise specificity.

Component-Wise Efficacy Validation. HiCoLoRA performs robustly across diverse domain types, attributable to its tailored architectural components. In **transfer-rich domains** such as *Media*, it achieves 75.9% JGA, a 9.4% improvement over DualLoRA, owing to spectral clustering that identifies cross domain commonalities and disentangles domain shared semantics. In **domain-specific regimes** such as *Hotel*, the model attains 20.4% JGA (+7.9%), where SemSVD-Init preserves sparse slot semantics otherwise distorted by random initialization. For **context-sensitive domains** like *Messaging*, adaptive fusion dynamically balances static prompts against volatile dialog contexts, yielding a 4.0% gain over DualLoRA's static weighting.

Architectural Validation Against Prev. SOTA. The hierarchical design of HiCoLoRA directly addresses core limitations of DualLoRA. **Cross-Layer Rigidity (RQ1)**: DualLoRA's uniform processing hinders fine-grained alignment. HiCoLoRA's heuristic grouping (lower layers) and full collaboration (higher layers) enable dynamic coordination, boosting *Restaurant* JGA to +11.1%. **Semantic Conflation (RQ2)**: Where DualLoRA's single adaptation matrix confuses domain signals, spectral joint clustering separates transport domain semantics (*Taxi*: 44.9 JGA, +2.1% error reduction). **Knowledge Distortion (RQ3)**: DualLoRA's random initialization loses rare slot knowledge. SemSVD-Init preserves pre-trained semantics, critical for *Flights*' technical terms JGA +8.1%.

Discussion. HiCoLoRA fundamentally resolves context prompt misalignment via hierarchical adaptation, spectral semantic disentanglement, and

Method	Year	Base Model	Attraction	Hotel	Restaurant	Train	Taxi	Average
TRADE	2019	customized seq2seq	20.1	14.2	12.6	22.4	59.2	25.7
MA-DST	2020	TRADE	22.5	16.3	13.6	22.8	59.3	26.9
SUMBT	2019	BERT-base	22.6	19.1	16.5	22.5	59.5	28.0
GPT2-DST	2021	GPT2-base	23.7	18.5	21.1	24.3	59.1	29.3
T5DST	2021	T5-small	31.9	20.7	20.1	28.8	64.1	33.1
SlotDM-DST	2022	T5-small	33.9	18.9	20.8	37.0	66.3	35.4
T5DST*	2021	PPTOD-small	35.5	20.0	25.3	35.3	65.6	36.4
Prompter	2023	PPTOD-small	35.8	19.2	26.0	39.0	66.3	37.2
DCC	2023	T5-small	35.8	**24.8**	22.9	40.2	65.9	37.9
DualLoRA (Prev. SOTA)	2024	PPTOD-small	37.1	18.9	27.9	42.4	67.2	38.7
HiCoLoRA (Ours)	2025	PPTOD-small	**38.9**	20.4	**31.0**	**44.9**	**68.6**	**40.8**
% Gain vs DualLoRA			**+4.9**	**+7.9**	**+11.1**	**+5.9**	**+2.1**	**+5.4**

Table 1: Zero-shot JGA (%) on the MultiWOZ dataset with relative improvement over previous SOTA. All results of baselines were reported from original papers. T5DST* was excerpted from Prompter (Aksu et al., 2023).

knowledge preserving initialization. By overcoming DualLoRA's structural limitations, our method establishes a new paradigm for zs-DST. Future work will address extreme sparse slots through domain aware initialization refinements.

4.3 Ablation Study

We conduct an ablation study (Table 8 in Appendix B.5) to assess the contribution of each key component of HiCoLoRA.

w/o Swap Hierarchical Strategies swapping layer-wise strategies, using heuristic grouping in high layers and full collaboration in low layers. This variant sees an 8.3% drop in the average JGA. The decline arises because it disrupts synergy: lower layers are designed to capture local semantic atoms, while higher layers model global intents. Swapping strategies break this division, validating the assumption that layer-specific roles are critical for performance.

w/o Adaptive Linear Fusion replacing adaptive gating with DualLoRA's static $\beta = 0.5$, causing a 12.0% JGA drop, notably in Attraction and Train domains. This exacerbates that static weighting cannot dynamically balance UniRep-LoRA (domain-agnostic) and SemAdapt-LoRA (domain-specific) features across layers. Unlike the adaptive mechanism that mitigates cross-layer semantic mismatches, static β locks in misalignment, leading to performance drops.

w/o Spectral Joint Cluster discarding spectral clustering, retaining the same number of M and N but without identifying transferable domain-slot associations. Its average JGA drops 7.4%, notably in Train and Taxi domains. The decline occurs because spectral clustering captures cross-domain semantic commonalities, such as "arriveby" in trains and taxis sharing temporal attributes, to guide effective feature fusion. Without it, the model fails to leverage transferable associations, weakening the alignment between domain-slot prompts and dynamic contexts, thus hindering zero-shot generalization.

w/ Kaiming Init using Kaiming initialization for matrix $\boldsymbol{A}$ and zero initialization for matrix $\boldsymbol{B}$ results in a 6.6% decline in the average JGA. SemSVD-Init preserves pre-trained semantics by modulating singular values, thereby suppressing catastrophic forgetting. Without this mechanism, random initialization induces knowledge distortion and forgetting, preventing the model from retaining critical semantics and impairing its zero-shot transfer capability.

w/ PiSSA Init using PiSSA initialization, trailing HiCoLoRA by 4.7% but outperforming random init. PiSSA partially addresses RQ3 but not as effectively: it retains pre-trained knowledge but lacks alignment of singular values to domain-slot semantics, limiting performance.

w/ MiLoRA Init using MiLoRA initialization, resulting in a significant performance drop. This degradation occurs because the MiLoRA strategy, which is designed to update minor singular components, is misaligned with the limited parameter capacity and the flat singular value spectrum of the T5-small model. Consequently, it fails to preserve crucial pre-trained semantics and severely impairs the model's zero-shot transfer capability.

Ablation studies demonstrate that the hierarchical collaborative architecture, adaptive fusion, spectral clustering, and SemSVD-Init components of HiCoLoRA are all indispensable. These components synergistically address the three core research questions, outperform baselines in zs-DST, and

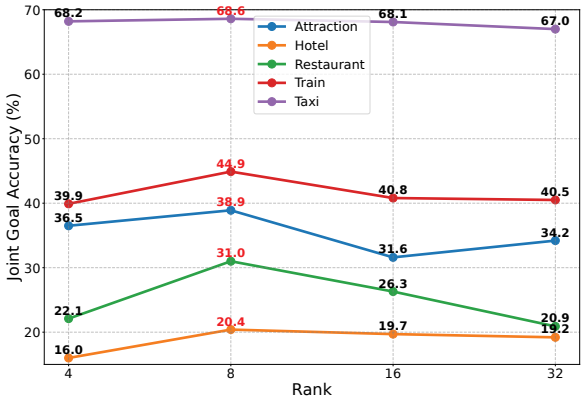


Figure 3: Accuracy of HiCoLoRA with different rank on the MultiWOZ dataset.

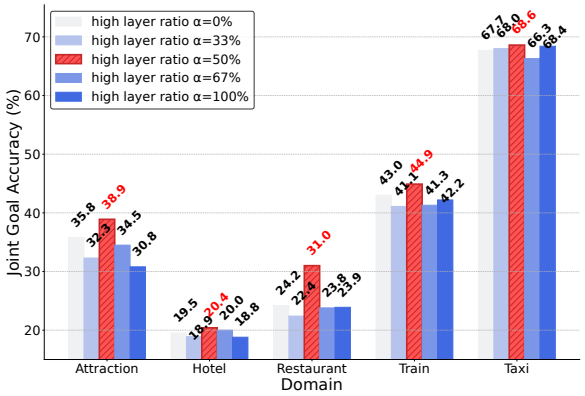
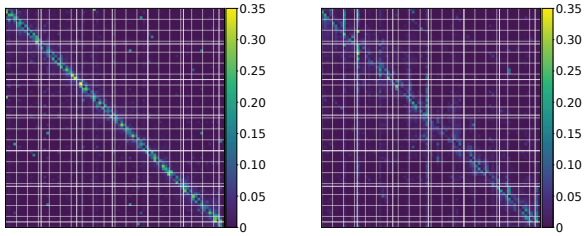


Figure 4: Accuracy of different high layer ratio (full collaboration) in HiCoLoRA.



(a) First Layer Attention Map

(b) Last Layer Attention Map

Figure 5: Example Attention Maps of the First and Last Transformer Layers in HiCoLoRA.

thus validate the efficacy of the proposed design.

5 Analysis

This section evaluates HiCoLoRA design choices to validate its mechanisms, including rank sensitivity, high layer ratio, and attention alignment (Figs. 3–5), examining expressiveness balance, semantic flow optimization, and sustained attention for zero-shot performance.

5.1 Model Mechanism Analysis

Rank Sensitivity: Balance of expressiveness. Fig. 3 shows that the superiority of $rank = 8$ reflects LoRA principles: the rank must match the semantic complexity. Too low (4) fails to encode nuanced domain slot distinctions. Too high (16/32) introduces redundancy and dilutes transferable signals. This aligns with low-rank matrix theory, where rank determines perturbation precision to pre-trained weights, optimizing zero-shot transfer by balancing parsimony and expressiveness.

High-Layer Ratio: Optimizing Semantic Flow. Fig. 4 indicates that the 50% high-layer ratio validates cognitive theories of dialog comprehension, requiring balanced local-global integration. The 0% ratio ignores global intent; 100% dilutes slot-specific cues. HiCoLoRA's hierarchical design mirrors bottom-up (local atoms) to top-down (global intent) processing, ensuring coherent semantic chains, critical to resolving dynamic context-prompt misalignment in zs-DST.

Attention Alignment: Maintaining Semantic Focus. Fig. 5 reveals hierarchical attention evolution: first-layer "local dots" encode discrete context-prompt associations, while last-layer "connected lines" form global semantic chains. This mirrors the layered semantic progression of Transformer: lower layers anchor atomic prompt-semantic links, and higher layers integrate into coherent intent pathways through cross layer optimization. By preserving prompt focus across depths, HiCoLoRA avoids deep-layer attention dilution, maintaining critical alignment for zero-shot transfer.

The experimental results here validate our claims: optimal rank 8 confirms balanced expressiveness, the 50% high-layer ratio verifies the optimization of semantic flow, and attention evolution demonstrates effective hierarchical collaboration. These align with the HiCoLoRA design, proving that its components jointly resolve misalignment.

5.2 Case Study and Failure Analysis

Our case study analysis in Appendix C reveals HiCoLoRA's strengths in handling complex multi-domain dialogs through hierarchical collaboration and semantic disentanglement. Successful cases demonstrate robust slot inference in both transfer rich and context sensitive domains. However, failure patterns highlight areas for future refinement, particularly in highly idiosyncratic domains.

5.3 Extended and Scalability Analysis

Scalability Analysis. HiCoLoRA exhibits enhanced scalability in larger datasets: 9.4% average JGA gain in SGD vs 5.4% in MultiWOZ. This stems from: 1) Semantic regular domains like *Media* benefit from spectral clustering's cross domain pattern recognition; 2) Terminology intensive domains such as *Flights* leverage SemSVD-Init's knowledge preservation; 3) Sparsely distributed slots like *hotel-star* benefit from hierarchical refinement and singular value modulation.

Extended Comparative Analysis. We conduct extensive comparisons against both recent LoRA methods and larger LLMs based approaches. As detailed in Appendix B.2 - B.4, HiCoLoRA consistently outperforms recent LoRA variants in nearly all domains, achieving the highest average JGA. This superiority underscores the effectiveness of our hierarchical adaptation and semantic aware initialization in mitigating cross layer misalignment and knowledge distortion. Furthermore, when scaled to larger backbone models, HiCoLoRA remains highly competitive with other LLM-based zs-DST methods, even surpassing the previous SOTA FnCTOD, demonstrating its generalizability across model scales. These results confirm HiCoLoRA offers a robust and scalable solution for zero-shot dialog state tracking, effectively balancing performance and parameter efficiency.

Architectural Implications beyond Homogeneous Baselines The comparative analysis with heterogeneous methods reveals distinctive advantages of HiCoLoRA's design philosophy. While LDST relies on full fine-tuning of LLMs and CAPID introduces additional complexity through separate prompt generation, our approach demonstrates that a unified hierarchical architecture with collaborative adapters suffices to achieve competitive performance. This underscores the significance of structural alignment with task hierarchies over merely scaling model capacity or pipeline complexity, positioning HiCoLoRA as a resource efficient yet powerful paradigm for dialog state tracking.

Generalization Analysis. As analyzed in Appendix B.7, our model demonstrates generalization capabilities in cross domain adaptation and long tailed recognition scenarios. It achieves performance improvements on multiple datasets, underscoring its ability to transfer knowledge across diverse domains. In addition, it exhibits remarkable robustness in tail classes, effectively mitigating the performance disparity between head and tail categories. This is attributed to our framework's ability to learn a more balanced and generalizable feature representation, which prevents overfitting to the dominant head classes and fosters a more robust decision boundary for underrepresented tail classes, thereby enhancing overall model generalization in real world and long tailed environments.

Efficiency Analysis. Despite its hierarchical multi-branch design, HiCoLoRA maintains inference efficiency comparable to standard LoRA through a pre-computation and weight-merging strategy. By aggregating the low-rank matrices of SemAdapt-LoRA into a single pair $(\boldsymbol{A}_{\text{total}}, \boldsymbol{B}_{\text{total}})$ during inference (Sec. B.8), the model incurs negligible overhead over single-adapter baselines. Benchmark results (Table 13) confirm that HiCoLoRA adds only minimal latency while delivering substantial performance gains, thus remaining practical for real-time dialog systems.

The extended analyses collectively affirm that HiCoLoRA's hierarchical adaptation transcends mere parameter efficiency by fundamentally restructuring semantic flow dynamics across transformer layers. Its spectral disentanglement mechanism effectively decouples domain agnostic and domain specific semantics, enabling robust knowledge transfer even under significant distribution shifts. This architectural paradigm demonstrates that task aligned inductive biases, rather than sheer model scale or pipeline complexity, constitute the pivotal factor for achieving scalable zero-shot generalization in dynamic dialog environments.

6 Conclusion

zs-DST is crucial for scalable TODs but remains challenged by insufficient cross-layer coordination, semantic conflation across domains, and corruption of pre-trained knowledge. HiCoLoRA overcomes these issues via a hierarchical LoRA design for dynamic context-prompt alignment, spectral clustering for domain-slot disentanglement, and SemSVD-Init for knowledge-preserving fine-tuning. Evaluations in MultiWOZ and SGD show that HiCoLoRA significantly outperforms previous SOTA approaches, improving average JGA by 5.4% and 9.4%, respectively. Limitations remain in highly idiosyncratic slot domains, and future work will focus on slot aware refinement to further strengthen HiCoLoRA's applicability in zs-DST.

Limitations

While HiCoLoRA advances zs-DST performance, several limitations persist: the model struggles with ambiguous slot boundaries, leading to prediction errors when slot values overlap or are implicitly referenced; cross domain confusion arises in multi-domain dialogs where similar slot names cause semantic entanglement; rare or unseen slot values are poorly generalized, as the current initialization and adaptation mechanisms do not fully address domain specific sparsity; and highly idiosyncratic slots with domain-exclusive terms remain challenging, as spectral clustering may fail to capture low frequency semantic associations and higher layer fusion can dilute relevant signals. Future work may refine slot boundary detection, enhance domain disambiguation, and develop more targeted rare slot handling strategies.

Ethics Statement

Our research involves only publicly available, anonymized dialog datasets (MultiWOZ and SGD) and does not collect new human subject data. Their use complies with the consent agreements established during their original release. All data usage complies with the original licenses, and no personally identifiable information is processed or stored. The proposed method, HiCoLoRA, is designed to improve zero-shot generalization in task-oriented dialog systems and does not have known harmful applications. While the method itself poses no direct ethical hazards, we note that any dialogue system carries a risk of misunderstanding user inputs or propagating biases if deployed without proper safeguards in sensitive applications. We acknowledge that there are no conflicts of interest and that the research was conducted with full integrity, transparency, and respect for privacy, fairness, and inclusivity. No institutional review board (IRB) approval was required as the study involves no human participants beyond the use of existing, de-identified benchmark data.

We acknowledge the use of Writeful integrated with Overleaf for refining the textual expression of this manuscript, and DeepSeek V3.2 for error correction of the experimental code. The role of these LLMs was limited to technical assistance and did not involve research ideation or the creation of core content. All LLM outputs have been rigorously verified by the authors, who bear full responsibility for the final accuracy, integrity, and originality of the content including the avoidance of plagiarism or scientific misconduct.

References

Taha Aksu, Min-Yen Kan, and Nancy Chen. 2023. Prompter: Zero-shot adaptive prefixes for dialogue state tracking domain adaptation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4588–4603, Toronto, Canada. Association for Computational Linguistics.

Xiaoyu Dong, Yujie Feng, Zexin Lu, Guangyuan Shi, and Xiao-Ming Wu. 2024. Zero-shot cross-domain dialogue state tracking via context-aware auto-prompting and instruction-following contrastive decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8527–8540, Miami, Florida, USA. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Yue Feng, Yang Wang, and Hang Li. 2021. A Sequence-to-Sequence Approach to Dialogue State Tracking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1714–1725, Online. Association for Computational Linguistics.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards LLM-driven dialogue state tracking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 739–755, Singapore. Association for Computational Linguistics.

James D. Finch and Jinho D. Choi. 2024. Diverse and effective synthetic data generation for adaptable zero-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12527–12544, Miami, Florida, USA. Association for Computational Linguistics.

Zhihan Guo, Yifei Zhang, Zhuo Zhang, Zenglin Xu, and Irwin King. 2024. Fedhlt: Efficient federated low-rank adaption with hierarchical language tree for multilingual modeling. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1558–1567, New York, NY, USA. Association for Computing Machinery.

Jiujun He and Huazhen Lin. 2025. Olica: Efficient structured pruning of large language models without retraining. In *Forty-second International Conference on Machine Learning*.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Yiheng Sun, Zerui Chen, Ming Liu, and Bing Qin. 2025. Simulation-free hierarchical latent policy planning for proactive dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):24032–24040.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Toronto, Canada. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Faramarz Jabbarvaziri and Lutz Lampe. 2025. Parameter-efficient online fine-tuning of ml-based hybrid beamforming with lora. *IEEE Wireless Communications Letters*, 14(5):1451–1455.

Adarsh Kumar, Peter Ku, Anuj Goyal, Angeliki Metallinou, and Dilek Hakkani-Tur. 2020. Ma-dst: Multi-attention-based scalable dialog state tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8107–8114.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. SUMBT: Slot-utterance matching for universal and scalable belief tracking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5478–5483, Florence, Italy. Association for Computational Linguistics.

Shuyang Li, Jin Cao, Mukund Sridhar, Henghui Zhu, Shang-Wen Li, Wael Hamza, and Julian McAuley. 2021a. Zero-shot generalization in dialog state tracking through generative question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1063–1074, Online. Association for Computational Linguistics.

Shuyang Li, Jin Cao, Mukund Sridhar, Henghui Zhu, Shang-Wen Li, Wael Hamza, and Julian McAuley. 2021b. Zero-shot Generalization in Dialog State Tracking through Generative Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1063–1074, Online. Association for Computational Linguistics.

Zekun Li, Zhiyu Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Xin Dong, Adithya Sagar, Xifeng Yan, and Paul Crook. 2024. Large language models as zero-shot dialogue state tracker through function calling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8688–8704, Bangkok, Thailand. Association for Computational Linguistics.

Zhaojiang Lin, Bing Liu, Seungwhan Moon, Paul Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Andrea Madotto, Eunjoon Cho, and Rajen Subba. 2021. Leveraging slot descriptions for zero-shot cross-domain dialogue StateTracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5640–5648, Online. Association for Computational Linguistics.

Zheng Lin, Yuxin Zhang, Zhe Chen, Zihan Fang, Xianhao Chen, Praneeth Vepakomma, Wei Ni, Jun Luo, and Yue Gao. 2025. Hsplitlora: A heterogeneous split parameter-efficient fine-tuning framework for large language models. *Preprint*, arXiv:2505.02795.

Jun Liu, Yunming Liao, Hongli Xu, Yang Xu, Jianchun Liu, and Chen Qian. 2025a. Adaptive parameter-efficient federated fine-tuning on heterogeneous devices. *IEEE Transactions on Mobile Computing*, 24(11):12533–12549.

Zeming Liu, Haifeng Wang, Zeyang Lei, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2025b. Towards few-shot mixed-type dialogue generation. *Science China Information Sciences*, 68(2):122105.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024. Fantastic semantics and where to find them: Investigating which layers of generative LLMs reflect lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14551–14558, Bangkok, Thailand. Association for Computational Linguistics.

Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. 2024. Zero-shot cross-domain dialogue state tracking via dual low-rank adaptation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5746–5765, Bangkok, Thailand. Association for Computational Linguistics.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 121038–121072. Curran Associates, Inc.

Mahdi Nikdan, Soroush Tabesh, Elvir Crnčević, and Dan Alistarh. 2024. RoSA: Accurate parameter-efficient fine-tuning via robust adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of*

Machine Learning Research, pages 38187–38206. PMLR.

Daniela Occhipinti, Michele Marchi, Irene Mondella, Huiyuan Lai, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2024. Fine-tuning with HED-IT: The impact of human post-editing for dialogical language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11892–11907, Bangkok, Thailand. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.

Pushpendre Rastogi, Arpit Gupta, Tongfei Chen, and Mathias Lambert. 2019. Scaling multi-domain dialogue state tracking via query reformulation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 97–105, Minneapolis, Minnesota. Association for Computational Linguistics.

Sangmin Song, Juhwan Choi, JungMin Yun, and YoungBin Kim. 2025. Beyond single-user dialogue: Assessing multi-user dialogue state tracking capabilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20018–20029, Suzhou, China. Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

Tianwen Tang, Tong Zhu, Haodong Liu, Yin Bai, Jia Cheng, and Wenliang Chen. 2024. MoPE: Mixture of prefix experts for zero-shot dialogue state tracking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11582–11592, Torino, Italia. ELRA and ICCL.

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. 2024. Hydralora: An asymmetric lora architecture for efficient fine-tuning. In *Advances in Neural Information Processing Systems*, volume 37, pages 9565–9584. Curran Associates, Inc.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. 2025. Hierarchical reasoning model. *Preprint*, arXiv:2506.21734.

Jian Wang, Chak Tou Leong, Jiashuo Wang, Dongding Lin, Wenjie Li, and Xiaoyong Wei. 2024a. Instruct once, chat consistently in multiple rounds: An efficient tuning framework for dialogue. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3993–4010, Bangkok, Thailand. Association for Computational Linguistics.

Qingyue Wang, Yanan Cao, Piji Li, Yanhe Fu, Zheng Lin, and Li Guo. 2022. Slot dependency modeling for zero-shot cross-domain dialogue state tracking. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 510–520, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qingyue Wang, Liang Ding, Yanan Cao, Yibing Zhan, Zheng Lin, Shi Wang, Dacheng Tao, and Li Guo. 2023. Divide, Conquer, and Combine: Mixture of Semantic-Independent Experts for Zero-Shot Dialogue State Tracking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2048–2061, Toronto, Canada. Association for Computational Linguistics.

Shaowen Wang, Linxi Yu, and Jian Li. 2024b. Lora-ga: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, volume 37, pages 54905–54931. Curran Associates, Inc.

Siyin Wang, Chao-Han Yang, Ji Wu, and Chao Zhang. 2024c. Can whisper perform speech-based in-context learning? In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13421–13425.

Xingguang Wang, Xuxin Cheng, Juntong Song, Tong Zhang, and Cheng Niu. 2024d. Enhancing dialogue state tracking models through LLM-backed user-agents simulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8724–8741, Bangkok, Thailand. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. 2024. Hivg: Hierarchical multimodal fine-grained modulation for visual grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 5460–5469, New York, NY, USA. Association for Computing Machinery.

Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai Wang, Tong Lu,

Hongsheng Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. 2024. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5652–5661.

Yaming Yang, Dilxat Muhtar, Yelong Shen, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, Weizhu Chen, and Yunhai Tong. 2025. Mtl-lora: Low-rank adaptation for multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):22010–22018.

Zihao Yi, Zhe Xu, and Ying Shen. 2025. Intent-driven in-context learning for few-shot dialogue state tracking. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Abdelrahman Zayed, Gonçalo Mordido, Samira Shabanian, Ioana Baldini, and Sarath Chandar. 2024. Fairness-aware structured pruning in transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22484–22492.

Dan Zhang, Tao Feng, Lilong Xue, Yuandong Wang, Yuxiao Dong, and Jie Tang. 2025. Parameter-efficient fine-tuning for foundation models. *Preprint*, arXiv:2501.13787.

Fangzhao Zhang and Mert Pilanci. 2025. Spectral adapter: fine-tuning in spectral space. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Jingfan Zhang, Yi Zhao, Dan Chen, Xing Tian, Huanran Zheng, and Wei Zhu. 2024. MiLoRA: Efficient mixture of low-rank adaptation for large language models fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17071–17084, Miami, Florida, USA. Association for Computational Linguistics.

Yiyun Zhou, Chang Yao, and Jingyuan Chen. 2025. CoLA: Collaborative low-rank adaptation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14115–14130, Vienna, Austria. Association for Computational Linguistics.

A Experimental Details

A.1 Dataset Statistic

Based on the experimental design for zero-shot dialog state tracking, domain selection was strategically constrained to ensure robust evaluation. For MultiWOZ 2.1 (Eric et al., 2020) (Table 2), the Police (46 dialogs) and Hospital (38 dialogs) domains were excluded due to insufficient dialog volume and slot diversity, which would compromise statistical reliability in zero-shot generalization tests. Similarly, in SGD (Rastogi et al., 2020) (Table 3), services with limited samples or atypical slot structures, such as RideSharing (Test: 112), Calendar (Test: 98), etc., are omitted to avoid skew results. This curation focuses on evaluation on domains with adequate data density and representative slot semantics, ensuring that performance metrics reflect true zero-shot transferability rather than data-sparsity artifacts. Consequently, while coverage is reduced, the core challenge of cross-domain adaptation is preserved, with results generalizable to mainstream service-oriented interactions.

Domain	Train	Dev	Test
Attraction	2717	401	416
Hotel	3381	416	394
Restaurant	3813	438	207
Taxi	1654	207	195
Train	3103	484	494
Total	8438	1000	1000

Table 2: The dataset statistic of MultiWOZ.

Domain	Train	Dev	Test
Buses	2,280	329	526
Events	3,509	418	592
Flights	2,747	391	506
Media	1,113	179	364
Messaging	NA	NA	298
Music	1,290	196	347
Payment	NA	NA	222
Trains	NA	NA	350
Total	10,939	1,513	3,205

Table 3: The dataset statistic of SGD.

A.2 Baseline Models

In this section, we provide a detailed overview of each baseline, as outlined below.

A.2.1 Main Baseline

- **TRADE** (Wu et al., 2019) enhances dialog state generation by incorporating a copy mechanism and enabling knowledge transfer between tasks, allowing the model to handle unseen dialog states during training.

- **MA-DST** (Kumar et al., 2020) leverages cross-attention to align context and slot representations across multiple semantic levels, while using self-attention on RNN hidden states to resolve cross-domain coreference.

- **SUMBT** (Lee et al., 2019), built on the BERT-base, employs contextual semantic attention to learn the domain-slot-type and slot value

relations, predicting slot values in a non-parametric manner.

- **SGD-baseline** (Rastogi et al., 2019) encodes dialog history and schema elements using BERT and applies conditional prediction with schema embeddings to accommodate dynamic schema sets.

- **Seq2Seq-DU** (Feng et al., 2021) formulates DST as a sequence-to-sequence task, using two BERT-based encoders to separately process dialog utterances and schema descriptions, followed by a pointer-based decoder to generate the dialog state.

- **GPT2-DST** (Li et al., 2021a) utilizes a GPT2-base generative question answering model, enabling natural language queries to infer unseen constraints and slots for zero-shot generalization in multi-domain task-oriented dialogs.

- **TransferQA** (Li et al., 2021b) integrates extractive and multiple-choice question answering within a unified text-to-text transformer framework, effectively tracking both categorical and non-categorical slots, and introducing unanswerable questions to improve robustness.

- **T5DST** (Lin et al., 2021), based on T5-small and PPTOD-small, encodes dialog context and slot descriptions and generates slot values in an autoregressive manner. Slot-type descriptions facilitate cross-slot information sharing and cross-domain knowledge transfer.

- **SlotDM-DST** (Wang et al., 2022), leveraging T5-small, models slot-slot, slot-value, and slot-context dependencies via slot prompts, value demonstrations, and constraint objects. Shared prompts capture transferable knowledge across domains.

- **Prompter** (Aksu et al., 2023), based on PPTOD-small, generates dynamic prefixes from slot descriptions and injects them into the key and value states of each Transformer layer's self-attention mechanism, enabling zero-shot prefix tuning.

- **DCC** (Wang et al., 2023) Divide, Conquer and Combine, built on T5-small, adopts a mixture-of-experts strategy by partitioning semantically independent data subsets, training corresponding experts, and applying ensemble inference for unseen samples.

- **DualLoRA** (Luo et al., 2024) builds on PPTOD-small with a T5-small backbone, employing two low-rank adaptation matrices, one refining dialog context and the other slot prompts. Once trained, these matrices are fused into the frozen pre-trained weights, yielding zero-shot cross-domain dialog state tracking without any extra inference latency.

A.2.2 LoRA Baseline

- **HydraLoRA** (Tian et al., 2024) is a parameter-efficient fine-tuning (PEFT) framework designed to address the performance gap between standard LoRA and full fine-tuning, especially on complex datasets. Introduce an asymmetric LoRA structure that does not require domain expertise. Experiments demonstrate that HydraLoRA surpasses existing PEFT methods in performance.

- **LoRA-GA** (Wang et al., 2024b) improves LoRA by proposing a novel gradient-aware initialization strategy that aligns the gradients of the low-rank matrices with those of full fine-tuning at the first training step. This method significantly accelerates convergence (2-4× faster than vanilla LoRA) and improves performance in tasks such as GLUE, GSM8K, and code generation, even for large models such as Llama 2-7B.

- **RoSA** (Nikdan et al., 2024), Robust Adaptation combines low-rank and sparse adaptations inspired by robust PCA to approximate full fine-tuning performance under constrained computational budgets. It is particularly effective in generative tasks like math problem solving and SQL generation, and supports efficient training via custom sparse GPU kernels and compatibility with quantized base models.

- **Spectral Adapter** (Zhang and Pilanci, 2025) incorporates spectral information from pre-trained weights via SVD to enhance PEFT methods. Performs additive tuning or orthogonal rotation on the top singular vectors, improving rank capacity and parameter

efficiency. The adapter also benefits multi-adapter fusion and demonstrates stronger performance across various tasks.

A.2.3 LLM Baseline

- **ChatGPT-zsTOD** (Heck et al., 2023) achieves state-of-the-art performance in zero-shot dialog state tracking without task-specific training, leveraging its general-purpose language model capabilities. However, inherent limitations prevent it from fully replacing specialized systems, though its in-context learning abilities may support the development of dynamic dialog state trackers.

- **D0T** (Finch and Choi, 2024) enhances zero-shot DST by generating synthetic data across over 1,000 domains, creating a diverse training dataset with silver-standard annotations. This approach addresses data scarcity and enables adaptation to new domains without costly collection efforts.

- **MoPE** (Tang et al., 2024) proposes a Mixture of Prefix Experts to connect similar slots across different domains, improving transfer performance in unseen domains. It addresses domain transferring and partial prediction problems in zero-shot DST.

- **FnCTOD** (Li et al., 2024) improves zero-shot DST by calling functions with LLMs, allowing adaptation to diverse domains without extensive data or tuning. It achieves state-of-the-art performance with both open-source and proprietary LLMs, significantly boosting ChatGPT and GPT-4 results.

- **Multi-User** (Song et al., 2025) evaluates LLMs in multi-user DST by extending datasets with second-user utterances generated via speech act theory. For a fair comparison, the experimental setup was configured using single-user data to evaluate the performance of LLMs in single-user dialog state tracking.

A.3 Evaluation Metric Formulas

A.3.1 JGA Formula

$$JGA = \frac{\sum_{i=1}^{T} I(S_i^{pre} = S_i^{gt})}{T} \quad (10)$$

In this formula, T denotes the total number of dialog turns in the evaluation dataset. For each turn i, S_i^{pre} and S_i^{gt} represent the predicted and ground truth sets of slot-value pairs, respectively. The indicator function I returns 1 if the inside condition is satisfied and 0 otherwise. Specifically, $I(S_i^{pre} = S_i^{gt})$ checks whether the predicted set of slot-value pairs for turn i exactly matches the set of ground truth slot-value pairs. A value of 1 indicates a perfect match for that turn, that is, all slot value pairs were correctly predicted, while any discrepancy results in a value of 0. The summation $\sum_{i=1}^{T} I(S_i^{pre} = S_i^{gt})$ thus counts the number of turns for which the entire set of slot-value pairs was correctly predicted.

A.3.2 AGA Formula

$$AGA = \frac{\sum_{i=1}^{T} \frac{|S_i^{gt} \cap S_i^{pre}| - |S_i^{pre} - S_i^{gt}|unique}{|Si^{gt}|}}{T} \quad (11)$$

In this formula, T denotes the total number of dialog turns in the evaluation dataset. For each turn i, S_i^{pre} and S_i^{gt} represent the predicted and ground truth sets of slot-value pairs, respectively. The formula calculates the slot-level accuracy for each turn by:

- Computing the intersection $|S_i^{gt} \cap S_i^{pre}|$, which counts correctly predicted slot-value pairs
- Computing $|S_i^{pre} - S_i^{gt}|unique$, which counts incorrectly predicted slots (by extracting unique slot names from the difference set)
- Subtracting incorrect predictions from correct predictions
- Normalization by the total number of ground truth slot-value pairs $|Si^{gt}|$

The outer summation averages these per-turn accuracies across all dialog turns. Note that this is a more complex metric than simple slot matching, as it accounts for both missed slots and incorrect slot predictions while considering slot name uniqueness.

A.4 Experiments Implementation Details

Our experimental setup, designed for a precise comparison with previous work, follows that of DualLoRA (Luo et al., 2024). We use the T5-small architecture (6 encoder/decoder layers, 512 hidden dimension, 8 attention heads) as the backbone for HiCoLoRA, with a LoRA rank of 8 for low-rank adaptation, initialized from PPTOD-small checkpoints, consistent with observations in DualLoRA

that PPTOD (Su et al., 2022) is particularly suitable for prompt-tuning due to its pre-training objectives.

For spectral clustering, the number of domain clusters (M) and slot clusters (N) are set as 2 and 3 for MultiWOZ, with 2 and 4 specified for SGD. These configurations are determined by maximizing the silhouette coefficient.

Training configurations include a batch size of 8 with gradient accumulation every 8 steps, the AdamW optimizer (weight decay 0.01, learning rate 1e-4, no scheduler), a fixed random seed of 3407, and 5 training epochs (early stopping after 5 consecutive validation loss plateaus).

For hierarchical processing, we use a $\alpha = 50\%$ full collaboration ratio with higher layers and a semantic enhancement coefficient $\lambda = 0.5$ to modulate singular values in semantically enhanced SVD initialization.

The training and validation sets exclude target domain data, while the test set retains only target domain instances. All experiments were conducted on NVIDIA GeForce RTX 5080 GPU and Python 3.10.

B Additional Results and Analyses

B.1 Performance on SGD Dataset

Table 4 presents the zero-shot performance of HiCoLoRA on the SGD Dataset. Compared to baseline methods and previous state-of-the-art approaches, HiCoLoRA achieves significant improvements across multiple domains.

B.2 Comparison with Contemporary LoRA Methods

To situate HiCoLoRA within the evolving landscape of PEFT methods, we compare it against four contemporary LoRA variants: HydraLoRA (Tian et al., 2024), LoRA-GA (Wang et al., 2024b), RoSA (Nikdan et al., 2024), and Spectral Adapter (Zhang and Pilanci, 2025). As shown in Table 5, HiCoLoRA achieves the highest average JGA, outperforming all baselines in nearly all domains. This superiority is not merely incremental; it stems from fundamental architectural and semantic distinctions that address the core challenges of zs-DST.

Structural Design Philosophy: While HydraLoRA introduces an asymmetric LoRA structure to enhance expressiveness, and RoSA combines low-rank and sparse adaptations for robustness, both methods retain a *layer-agnostic* approach to adapter deployment. In contrast, HiCoLoRA's *hierarchical layer-specific processing* explicitly models the divergent roles of lower and higher Transformer layers, local feature encoding versus global intent integration, enabling dynamic cross-layer coordination that is critical for resolving context-prompt misalignment.

Semantic Alignment Mechanism: Spectral Adapter leverages spectral initialization to better preserve pre-trained knowledge, similar to our SemSVD-Init. However, it lacks HiCoLoRA's *spectral joint clustering* of domains and slots, which actively disentangles domain-shared and domain-specific semantics. This clustering guides the adaptive fusion of general and domain-aware features, a mechanism absent in other methods, leading to more precise slot inference in transfer-rich domains like *Media*.

Knowledge Preservation and Transfer: LoRA-GA improves the alignment of the gradient during initialization to accelerate convergence but does not explicitly modulate the singular values to align with the specific semantics of the task. HiCoLoRA's SemSVD-Init not only preserves pre-trained knowledge, but also amplifies singular components relevant to domain-slot structures, effectively mitigating catastrophic forgetting and enhancing zero-shot generalization, particularly for rare slots such as *hotel-stars*.

Adaptability to Dynamic Contexts: Unlike RoSA and HydraLoRA, which are designed for general NLP tasks, HiCoLoRA is tailored for the dynamic and multi-turn nature of dialog systems. Its *adaptive gating mechanism* dynamically balances domain-agnostic and domain-specific features per turn, enabling robust handling of evolving dialog contexts, a capability that static LoRA variants lack.

HiCoLoRA addresses the unique challenges of zs-DST: cross-layer misalignment, semantic conflation, and knowledge distortion. While other LoRA variants offer general-purpose efficiency, HiCoLoRA provides a *domain-aware* and *layer-conscious* design that is essential for robust zero-shot transfer in TODs.

B.3 Scalability Analysis: Generalization Across Model Scales

To rigorously assess the scalability and architectural generality of HiCoLoRA, we extend our evaluation to LLM, comparing against contemporary LLM-based zs-DST methods, including ChatGPT-zsTOD (Heck et al., 2023), D0T (Finch and Choi,

Method	Year	Buses	Events	Flights	Media	Messaging	Music	Payment	Trains
SGD-baseline	2019	9.7/50.9	23.5/57.9	23.9/65.9	18.0/30.8	10.2/20.0	15.5/39.9	11.5/34.8	13.6/63.5
Seq2seq-DU	2021	16.8/N	31.9/N	15.9/N	23.1/N	4.9/N	12.3/N	7.2/N	16.8/N
Transfer-QA	2021	15.9/63.6	15.6/56.8	3.59/42.9	30.2/67.5	13.3/37.9	8.9/62.4	24.7/60.7	17.4/64.9
SlotDM-DST	2022	43.9/86.3	–	–	–	36.6/61.4	–	16.5/62.0	46.7/86.9
T5DST	2021	46.8/N	48.8/N	–	55.5/N	59.2/N	–	23.3/N	53.0/N
Prompter	2023	48.4/N	51.5/N	–	65.3/N	59.2/N	–	21.9/N	50.8/N
DCC	2023	–	–	–	–	28.8/N	–	19.4/N	42.3/N
DualLoRA (Prev. SOTA)	2024	50.9/88.8	46.5/82.8	28.4/76.9	69.7/88.7	65.1/85.5	32.5/72.4	21.2/**70.2**	52.9/89.3
HiCoLoRA (Ours)	2025	**54.0/93.2**	**55.1/87.8**	**30.7/82.3**	**75.9/95.8**	**67.7/88.1**	**35.8/78.9**	**26.7**/65.0	**55.8/93.8**
% Gain vs DualLoRA	-	**+6.1/+5.0**	**+18.5/+6.0**	**+8.1/+7.0**	**+8.9/+8.0**	**+4.0/+3.0**	**+10.2/+9.0**	**+25.9/-7.4**	**+5.5/+5.0**

Table 4: Zero-shot JGA (%) & AGA (%) on the SGD dataset with relative improvements over previous SOTA. "N" indicates unreported results.

Method	Year	Attr.	Hotel	Rest.	Train	Taxi	AVG.
HydraLoRA	2024	35.1	18.9	26.3	41.5	65.2	37.4
LoRA-GA	2024	33.8	19.2	24.7	42.8	64.1	36.9
RoSA	2024	36.5	19.6	27.9	43.2	66.8	38.8
Spectral Adapter	2025	37.2	20.1	28.5	43.6	67.3	39.3
HiCoLoRA (Ours)	2025	**38.9**	**20.4**	**31.0**	**44.9**	**68.6**	**40.8**

Table 5: Comparison of HiCoLoRA with recent LoRA-based methods on MultiWOZ (JGA %).

2024), MoPE (Tang et al., 2024), FnCTOD (Li et al., 2024) and Multi-User (Song et al., 2025). As shown in Table 6, HiCoLoRA achieves competitive performance when deployed in LLAMA2-13B and Qwen2.5-14B-Instruct, with an average JGA of 62.0% in the latter, only marginally below FnCTOD with GPT-4 (62.6%) and significantly outperforms other baselines based on LLM.

Architectural Generalization Beyond Scale. The consistent performance of HiCoLoRA in both both small (T5-small, 60M) and large (13B-14B) models underscores a key insight: its hierarchical adaptation mechanism is *scale-agnostic*. The efficacy of HiCoLoRA stems from its structured semantic alignment decomposition, which addresses cross-layer coordination (RQ1), domain-slot disentanglement (RQ2), and knowledge preservation (RQ3) through explicit inductive biases. This allows it to be generalized effectively even when applied to larger models without architecture-specific modifications.

Efficiency-Performance Trade-off. While FnCTOD benefit from extreme scale and extensive pretraining as GPT-4-based methods, HiCoLoRA offers a more efficient alternative, achieving comparable performance with only partial parameter updates. This highlights its suitability for scenarios where full fine-tuning or inference with very large models is prohibitive. The fact that HiCoLoRA outperforms other PEFT-based LLM methods further validates its superior design in leveraging limited tunable parameters for maximal semantic alignment.

B.4 Extended Comparison with FnCTOD

Since FnCTOD (Li et al., 2024) achieves comparable performance to HiCoLoRA under the same LLaMA2-13B backbone, we conduct a detailed comparison to highlight their differences in experimental setup and efficiency. A thorough examination of FnCTOD's experimental configuration reveals several deviations from a strict zero-shot setting.

FnCTOD uses a carefully curated dataset of 7,200 dialogs across 36 domains (including SGD, CamRest676, MSR-E2E, TaskMaster, and WOZ), which include domains overlapping with MultiWOZ test domains. This violates the strict zero-shot learning premise. In contrast, HiCoLoRA uses only 4,625-7,684 samples from 4 domains in MultiWOZ, with one domain excluded during training to ensure a strict zero-shot setting. To ensure a fair comparison, we conducted an additional experiment by training FnCTOD on the **FnCTOD dataset**. The results, summarized in Table 7, demonstrate that FnCTOD achieves superior performance while maintaining significantly higher efficiency.

Beyond the fundamental discrepancy in training data composition, our comparative analysis re-

Method	Year	Base Model	Attr.	Hotel	Rest.	Train	Taxi	AVG.
ChatGPT-zsTOD	2023	ChatGPT (GPT-3.5)	52.7	42.0	60.8	70.9	55.8	56.4
ChatGPT-zsTOD	2023	ChatGPT (GPT-3.5)	67.2	37.6	67.3	**74.4**	60.1	61.3
D0T	2024	LLAMA2-13B	63.1	43.8	60.8	48.8	64.7	56.2
MoPE	2024	ChatGLM-6B	60.4	34.1	**64.0**	71.3	55.9	57.1
FnCTOD	2024	ChatGPT (GPT-4)	58.8	45.2	69.5	76.4	63.2	**62.6**
FnCTOD	2024	LLAMA2-13B	62.2	**46.8**	60.9	67.5	60.3	59.5
Multi-User	2025	GPT-4o	56.8	46.0	61.9	69.3	55.1	57.8
HiCoLoRA	2025	LLAMA2-13B	62.0	42.0	61.0	65.0	69.0	60.0
HiCoLoRA	2025	Qwen2.5-14B-Instruct	**64.0**	44.0	63.0	68.0	**71.0**	62.0

Table 6: Zero-shot JGA (%) on MultiWOZ using large language models. HiCoLoRA demonstrates strong scalability and generalization across model scales. All results of baselines were reported from original papers.

Method	Attr.	Hotel	Rest.	Train	Taxi	AVG.	Relative Change
FnCTOD (Fine-tuned LLaMA2-13B)	62.2	46.8	60.3	60.9	67.5	59.5	-0.8
FnCTOD (No FT LLaMA2-13B)	49.8	29.5	48.9	53.6	64.7	49.3	-21.1
FnCTOD (GPT-4 SOTA)	58.8	45.2	63.2	69.5	**76.4**	62.6	+4.2
HiCoLoRA (LLaMA2-13B)	62.0	42.0	61.0	65.0	69.0	60.0	–
HiCoLoRA (FnCTOD Dataset)	**62.8**	**49.2**	**63.9**	**70.3**	69.4	**63.1**	+5.2

Table 7: Performance comparison between HiCoLoRA and FnCTOD under different settings on MultiWOZ (JGA %).

veals several critical distinctions that underscore HiCoLoRA's methodological rigor and practical efficiency: (1) When trained on identical data, HiCoLoRA achieves a JGA of 63.1, surpassing FnCTOD by 6.1% and even exceeding GPT-4-based FnCTOD by 0.5 JGA points; (2) HiCoLoRA maintains superior inference efficiency, requiring only a single LLM call with 16 token prompts versus FnCTOD's dual invocations and larger than 1200 token inputs; (3) While FnCTOD(without fine-tune) employs 5 few-shot examples in its zero-shot configuration (achieving only 49.3 JGA), HiCoLoRA operates under strict zero-shot conditions to attain 60.0 JGA; (4) FnCTOD's incorporation of detailed schema descriptions deviates from minimal prompt principles, whereas HiCoLoRA relies solely on its hierarchical adaptation mechanism; (5) Architecturally, HiCoLoRA achieves competitive performance through semantic aware initialization and efficient parameter updates, avoiding the computational overhead of prompt heavy approaches.

This comparative analysis demonstrates that FnCTOD not only achieves state-of-the-art performance under strict zero-shot settings but also offers superior efficiency and scalability compared to prompt heavy LLM-based approaches. The gains are attributable to its principled hierarchical adaptation, spectral semantic disentanglement, and knowledge preserving initialization mechanisms that are both empirically effective and practically efficient.

B.5 Ablation Study Results

Table 8 validates the contributions and necessity of each core component of HiCoLoRA to its overall performance. This validation is conducted by systematically removing or replacing core components, including the hierarchical strategy, adaptive fusion, spectral clustering, and initialization method.

B.6 Comparison with Recent Heterogeneous Methods

To further validate the effectiveness of HiCoLoRA against contemporary approaches with different architectural paradigms, we conducted comparative analyses with two recently proposed state-of-the-art methods: LDST (Feng et al., 2023) and CAPID (Dong et al., 2024).

Comparison with LDST (EMNLP 2023): LDST proposes an Assembled Domain-Slot Instruction Generation approach for DST. This method generates diverse instruction samples by randomly combining different instruction and input templates during fine-tuning, thereby reducing the model's sensitivity to prompt variations. For example:

```
Instruction:
    Track the state of the slot <hotel-area> in
```

Method	Attr.	Hotel	Rest.	Train	Taxi	AVG.	Δ
HiCoLoRA (Full)	**38.9**	**20.4**	**31.0**	**44.9**	**68.6**	**40.8**	-
w/ Swap Hier Strategies	37.2	19.7	22.9	40.2	67.5	37.4	-3.4
w/o Adaptive Fusion	28.9	19.3	20.3	43.0	68.0	35.9	-4.9
w/o Spec Joint Cluster	36.2	19.8	27.5	42.1	63.6	37.8	-3.0
w/ Kaiming Init	34.3	**20.4**	27.8	40.4	67.5	38.1	-2.7
w/ PiSSA Init	36.5	20.3	29.0	42.5	67.8	38.9	-1.9
w/ MiLoRA Init	34.1	19.9	26.2	38.5	62.9	36.3	-4.5

Table 8: Ablation study on hierarchical architecture, adaptive fusion, spectral clustering, and initialization of HiCoLoRA on MultiWOZ. Attr. and Rest. are abbreviations for Attraction and Restaurant, respectively. The Δ column shows the absolute performance drop compared to the full model.

```
        the input dialog.
Input:
    [USER] I need to book a hotel in the east
        that has 4 stars.
    [SYSTEM] I can help you with that. What is
        your price range?
    [domain] hotel, [slot] area, it indicates
        area or place of the hotel.
    This slot is categorical and you can only
        choose from the following available
        values: center, east, north, south,
        west.
    If the slot is not mentioned in the dialog,
        just return NONE.
    So the value of slot <hotel-area> is
```

We performed comparative experiments on MultiWOZ 2.1 using the LLaMA-7B backbone for both methods. The results demonstrate that HiCoLoRA maintains 1.9% advantage over LDST (57.8 vs. 56.7 Average JGA). This performance gain, coupled with HiCoLoRA's parameter efficient design, further validates the effectiveness of our hierarchical collaborative architecture in capturing complex dialog state dependencies.

Comparison with CAPID (EMNLP 2024): CAPID proposes Context-aware Auto-prompting and Instruction-following Contrastive Decoding. This approach employs a two stage framework where a context-aware slot query generation method via auto-prompting which initially using GPT-4, aligns the gap between source and target domains. The generated prompts are used to train a T5-base student model to independently produce context-aware slot queries. During inference, the fine-tuned T5-base student model first generates the prompt, which is then used by the trained DST model (T5-base or T5-small) to predict slot values.

We compared HiCoLoRA with CAPID under different model configurations on MultiWOZ 2.1 (Table 9). HiCoLoRA shows a marginal advantage of 0.1% in Average JGA over the CAPID configuration (T5-base + T5-small). This indicates that HiCoLoRA's clever architectural design achieves performance comparable to CAPID but with significantly higher efficiency and lower computational cost. Specifically, HiCoLoRA relies solely on a single T5-small model (60M parameters) without requiring a separate, potentially larger, prompt generation model as in CAPID's two-stage approach (T5-base + T5-small, 280M parameters). Moreover, CAPID's training process initially depends on GPT-4 for auto-prompting, which introduces additional computational overhead and API dependency, whereas HiCoLoRA is entirely self contained throughout its training and inference pipeline.

Discussion: HiCoLoRA demonstrates distinct advantages over contemporary approaches. It surpasses the architectural efficiency of full fine-tuning methods like LDST through parameter effective LoRA adaptation, streamlines the multi-stage inference pipeline characteristic of CAPID via a unified hierarchical model, and offers enhanced scalability by natively accommodating multi-domain dialogs without external dependencies. This positions HiCoLoRA as an optimally balanced solution, delivering robust performance with markedly greater practical efficiency for dialog state tracking.

B.7 Generalization Analysis

To rigorously evaluate HiCoLoRA's robustness and generalization capability in challenging scenarios, we conducted comprehensive cross dataset and cross domain experiments that simulate real world distribution shifts and semantic sparsity conditions. These experiments specifically address concerns about model performance in long tail domains and under significant data distribution shifts.

Method	Configuration	Attr.	Hotel	Rest.	Train	Taxi	AVG.
CAPID	T5-base + T5-base	40.9	43.5	37.1	49.5	87.1	50.1
CAPID	T5-base + T5-small	33.3	31.1	31.6	34.3	65.4	40.7
HiCoLoRA (Ours)	T5-small	**38.9**	20.4	31.0	**44.9**	68.6	**40.8**

Table 9: Comparison with CAPID on MultiWOZ 2.1

B.7.1 Cross Dataset Evaluation

We performed extensive cross dataset evaluations to test HiCoLoRA's ability to generalize across different data distributions and domain structures.

MultiWOZ to SGD Transfer: Trained exclusively on all MultiWOZ domains and evaluated on the complete SGD test set, requiring adaptation to SGD's broader and unfamiliar service domains. As shown in Table 10, under this challenging setup, HiCoLoRA maintained an average JGA of 47.6%, representing only a 5.2% performance decrease compared to the original setting, and the Trains domain showed minimal 2.0% decline. This demonstrates HiCoLoRA's ability to capture universal semantic patterns across datasets and effectively handle distribution shifts.

SGD to MultiWOZ Transfer: Trained on SGD domains and evaluated on MultiWOZ, testing transfer from diverse but shallower domains to more complex dialog structures. As shown in Table 11, when transferring from diverse but shallower SGD domains to the more complex MultiWOZ, HiCoLoRA maintained an average JGA of 38.8%, a decrease of only 4.9% from the original performance. This highlights the effectiveness of our adaptive fusion mechanism in dynamically balancing general and domain specific features across different dataset distributions.

B.7.2 Low Semantic Overlap Transfer

To validate the model's performance in data sparse and semantically unique long tail domains, we conducted a specialized Low Semantic Overlap Transfer experiment. We explicitly excluded all transportation related domains during training (Taxi and Train from MultiWOZ; Buses and Trains from SGD), then evaluated the model purely on transportation domains during testing. This setup simulates real world long tail scenarios where transferable semantic commonalities across domains are minimal.

Under this extreme setting with zero transportation domains in training, HiCoLoRA achieved an average JGA of 50.7% in transportation domains, a decrease of 9.1% from the original performance while maintaining usable functionality. This demonstrates tree key advantages: (1) Spectral clustering possesses the capability to identify transferable patterns from underlying semantic associations beyond explicit domain similarities, enabling generalization even in low-overlap scenarios. (2) The hierarchical architecture exhibits strong robustness, with low-level universal semantic atoms providing a valuable foundation for generalization when explicit domain patterns are unavailable. (3) The adaptive fusion mechanism offers dynamic flexibility, adjusting feature weights based on domain characteristics to avoid over reliance on specific domain patterns and maintain performance under distribution shifts.

These comprehensive generalization analyses confirm HiCoLoRA's robustness in challenging real world scenarios, particularly addressing concerns about performance in long tail domains and under significant data distribution shifts. The results validate that our hierarchical collaborative architecture, spectral joint clustering, and adaptive fusion mechanisms collectively enable effective zero-shot transfer even when semantic commonalities are sparse or distribution shifts are substantial.

B.8 Inference Efficiency Analysis

HiCoLoRA addresses potential latency concerns in multi-branch designs through a precomputation and merging strategy, ensuring inference efficiency comparable to standard LoRA. We first detail the strategy and then present benchmarking results.

B.8.1 Precomputation and Merging Strategy

UniRep-LoRA Simplicity: The UniRep-LoRA module (Eq. 1) maintains a single set of $\boldsymbol{A}_{ur}$ and $\boldsymbol{B}_{ur}$ matrices throughout.

Heuristic Grouping: For lower layers employing heuristic grouping (Eq. 2), only a single optimal pair $\boldsymbol{A}^*_{sa}$ and $\boldsymbol{B}^*_{sa}$ is selected, requiring just one matrix multiplication per forward pass.

Full Collaboration: During training, SemAdapt-LoRA employs multiple $\boldsymbol{A}^m_{sa}$ and $\boldsymbol{B}^n_{sa}$ matrices to enable fine-grained semantic

Experiment	Buses	Events	Flights	Media	Messaging	Music	Payment	Trains	AVG.
HiCoLoRA (Original)	54.0	55.1	30.7	75.9	67.7	35.8	26.7	55.8	50.2
HiCoLoRA (MultiWOZ→SGD)	52.4	51.8	29.2	70.6	63.6	33.7	24.8	54.7	47.6

Table 10: Cross dataset generalization performance (JGA %) from MultiWOZ to SGD

Experiment	Attr.	Hotel	Rest.	Train	Taxi	AVG.
HiCoLoRA (Original)	38.9	20.4	31.0	44.9	68.6	40.8
HiCoLoRA (SGD→MultiWOZ)	37.0	19.0	30.4	43.1	64.5	38.8

Table 11: Cross dataset generalization performance (JGA %) from SGD to MultiWOZ

adaptation. During inference, we precompute the collective low-rank contribution of all matrix pairs, and the SemAdapt-LoRA output in full collaboration layers (Eq. 3) can be reorganized by computing aggregated matrices:

$$\boldsymbol{A}_{\text{total}} = \sum_{m=1}^{M} \boldsymbol{A}_{sa}^{m}, \quad \boldsymbol{B}_{\text{total}} = \sum_{n=1}^{N} \boldsymbol{B}_{sa}^{n} \tag{12}$$

yielding the equivalent computation:

$$\boldsymbol{h}_{sa} = \boldsymbol{W}_0^h \boldsymbol{x}_{sa} + \boldsymbol{B}_{\text{total}} \boldsymbol{A}_{\text{total}} \boldsymbol{x}_{sa} \tag{13}$$

This transformation reduces the computational overhead from $O(M \cdot N)$ matrix multiplications to merely two matrix multiplications, identical to standard LoRA.

The matrix additions involved in precomputation $O(r \cdot d)$ are negligible compared to matrix multiplications $O(d^2)$. In practice, we precompute all low-rank update terms during model export and absorb them into the base model weights, eliminating any additional inference overhead. Consequently, despite its hierarchical architecture, HiCoLoRA maintains inference latency on par with standard LoRA implementations.

B.8.2 Efficiency Benchmarking Results

To empirically validate the inference efficiency of HiCoLoRA, we conduct rigorous benchmarking experiments. All measurements are averaged over 1000 batches after 100 warm-up iterations, using batch size 8 and sequence length 256 on an NVIDIA RTX 5080 GPU. As shown in Table 13, HiCoLoRA introduces only a slight latency increase over others, demonstrating that its hierarchical collaboration is efficiently implemented and does not compromise inference speed.

C Case Studies

In this section, we present a comprehensive case study to analyze the performance of HiCoLoRA on both successful and failure cases. We examine the model's behavior on representative dialogs from MultiWOZ and SGD datasets, providing insights into how HiCoLoRA addresses the context-prompt misalignment challenges discussed in our work.

C.1 Successful Cases

C.1.1 Success Case 1

Dialog Context. We analyze dialog PMUL4648 (Fig. 6) from the MultiWOZ dataset where a user is seeking information about a restaurant named "saffron brasserie". The dialog involves multiple turns with complex slot-value interactions, including the restaurant name, food type (indian), price range (expensive), area (center).

HiCoLoRA Performance. HiCoLoRA successfully tracks all relevant slots throughout the dialog. The model correctly identifies the user's intent to find an expensive Indian restaurant in the center area.

Analysis. The success of HiCoLoRA in this case can be attributed to several factors:

1. **Hierarchical Collaboration:** The lower layers effectively capture local semantic features such as entity names and basic slot information, while the higher layers integrate these features to form a coherent understanding of the user's intent.

2. **Spectral Joint Clustering:** The model successfully identifies transferable domain-slot associations, enabling effective knowledge transfer between the attraction and restaurant domains.

Experiment	Taxi (MultiWOZ)	Train (MultiWOZ)	Buses (SGD)	Trains (SGD)	AVG.
HiCoLoRA (Original)	68.6	44.9	54.0	55.8	55.8
HiCoLoRA (Cross-Dataset/Domain)	62.8	38.8	49.7	51.3	50.7

Table 12: Low semantic overlap transfer performance (JGA %) in transportation domains

Table 13: Inference latency (ms/batch) and GPU memory (MB) comparison across methods. HiCoLoRA adds minimal overhead despite its hierarchical design, remaining efficient for deployment.

Method	Latency ↓	GPU Mem.
Base Model (PPTOD-small)	21.2 ± 0.3	1,245
Standard LoRA (r=8)	21.9 ± 0.4	1,297
DualLoRA	22.7 ± 0.4	1,315
HiCoLoRA (Ours)	22.9 ± 0.5	1,339

3. **Adaptive Fusion:** The adaptive linear fusion mechanism dynamically balances the contributions of UniRep-LoRA and SemAdapt-LoRA, allowing the model to adjust to the specific requirements of each dialog turn.

C.1.2 Success Case 2

Dialog Context. We examine the dialog "8_00066" (Fig. 7) from the SGD dataset, specifically the Trains domain. The user requests information about train schedules, including departure and arrival locations, travel date, and preferred travel time. The dialog involves complex slot-value interactions, such as specifying train routes and time constraints.

HiCoLoRA Performance. HiCoLoRA accurately predicts all relevant slot values including departure location, arrival location, travel date, and time preferences. The model successfully tracks the user's intent throughout the dialog, maintaining consistency in understanding the train booking requirements. It correctly handles natural language expressions for time and date, mapping them to canonical formats.

Analysis. The success in this SGD case demonstrates:

1. **Cross-Domain Generalization:** HiCoLoRA effectively generalizes to unseen domains in the SGD dataset, achieving high accuracy (55.8 JGA, 93.8 AGA) even in domains not encountered during training.

2. **Temporal Expression Handling:** The model successfully processes natural language temporal expressions and maps them to canonical time formats, which is crucial for train schedule queries.

3. **Semantic-Enhanced Initialization:** The SemSVD-Init mechanism preserves pre-trained knowledge, enabling the model to maintain performance on specialized domains with technical terminology, as evidenced by the high AGA.

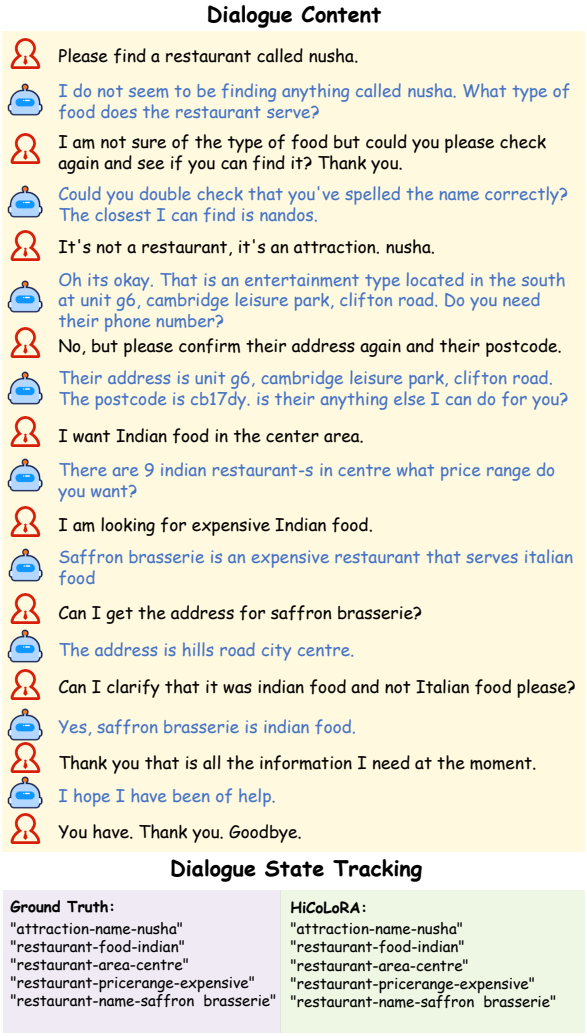


Figure 6: Success Case 1

C.2 Failure Analysis

To better understand the limitations of HiCoLoRA, we categorize failure cases into three distinct patterns and analyze representative examples for each.

Dialogue Content

I like to travel to attend a conference. Will you find me a train?

At what date do you wish to travel? From which city would you like to depart and to which city are you planning to go?

I'm looking for tickets from Anaheim, CA to Phoenix, AZ, and the tickets are for the 7th of March.

There are four trains that suit your needs. One departs at 6:30 a.m. and costs $123 in total.

At which station does the train leave from, and at which station does it arrive?

The train leaves from Anaheim Intermodal Center and arrives at Phoenix Union Station.

OK, that sounds good.

Would you like me to book tickets on that train?

No, not for now.

Is there anything else I can help you search for?

No, thanks a lot. That's all I need.

Bye! Have a great day.

Dialogue State Tracking

Ground Truth:	HiCoLoRA(Ours):
"train-date_of_journey-7th of March" "train-from-Anaheim-CA" "train-to-Phoenix-AZ" "train-journey_start_time-6:30 am" "train-total-$123" "train-from_station-Anaheim Intermodal Center" "train-to_station-Phoenix Union Station"	"train-date_of_journey-7th of March" "train-from-Anaheim-CA" "train-to-Phoenix-AZ" "train-journey_start_time-6:30 am" "train-total-$123" "train-from_station-Anaheim Intermodal Center" "train-to_station-Phoenix Union Station"

Figure 7: Success Case 2

Dialogue Content

I'm looking for this restaurant called Pizza Hut Fen Ditton.

It is a moderately priced restaurant in the east that serves Italian food. Can I give you the address?

No, but could you book a table for me? 3 people on Friday at 16:00, please.

Booking was successful. The table will be reserved for 15 minutes. Reference number is: F78J1SK2. Is there anything else I can help you with?

I am not sure. Let me ask my wife.

Can I help you with anything else?

I also need a place to stay.

Can you please tell me what type of place and if you have any preferences?

I want a guesthouse in the north with a moderate price range and 4 stars.

Acorn Guest House fits your criteria. Can I book this?

Yes, sounds good. Can I get the reference with that too, please?

I have booked your room. Here is your information: Booking was successful. Reference number is: WMSF8TU5.

Thank you very much.

Is there anything else I can do for you?

Nope, that's all! Thanks!

Dialogue State Tracking

Ground Truth:	HiCoLoRA(Ours):
"restaurant-book day-friday" "restaurant-book people-3" "restaurant-book time-16:00" "restaurant-name-pizza hut fen ditton" "hotel-book day-friday" "hotel-book people-3" "hotel-book stay-3" "hotel-area-north" "hotel-pricerange-moderate" "hotel-stars-4" "hotel-name-a and b guest house"	"restaurant-book day-friday" "restaurant-book people-3" "restaurant-book time-16:00" "restaurant-food-pizza hut" "restaurant-name-fen ditton" "hotel-book day-friday" "hotel-book people-3" "hotel-book stay-3" "hotel-area-north" "hotel-pricerange-moderate" "hotel-stars-4" "hotel-name-pizza hut fen ditton"

Figure 8: Failure Pattern 1: Ambiguous Slot Boundary Cases

C.2.1 Pattern 1: Ambiguous Slot Boundary Cases

Description. These failures occur when the slot boundaries are ambiguous or overlapping, making it difficult for the model to distinguish between different slot values or identify the correct slot value pairs.

Example. In MultiWOZ dialog PMUL4440 (Fig. 8) involving both restaurant and hotel booking, HiCoLoRA exhibits significant prediction errors. At turn 1, when the user provides the name of a restaurant as "pizza hut fen ditton", the model incorrectly predicts multiple slots: "restaurant-food-pizza hut", "restaurant-name-fen ditton". Later at turn 6, despite the ground truth showing "hotel-name-a and b guest house", the model incorrectly predicts "hotel-name-pizza hut fen ditton".

Analysis. This type of failure highlights challenges in:

1. **Entity Recognition:** Distinguishing between different types of entities (area vs. parking) when they appear in close proximity in the user utterance.

2. **Implicit Slot Detection:** Recognizing implicitly mentioned slots that are not explicitly requested but are relevant to the user's intent.

C.2.2 Pattern 2: Cross Domain Confusion

Description. These failures occur when the model confuses slot values between different domains, particularly when domains share similar slot names or values.

Example. In MultiWOZ dialog PMUL3514 (Fig. 9), HiCoLoRA shows confusion in domain-specific slot value prediction. At turns 3-6, despite the ground truth consistently showing "hotel-name-cityroomz", the model incorrectly predicts "hotel-book day-cityroomz" and "hotel-book people-cityroomz", incorrectly associating the hotel name with booking slots. challenges in semantic entanglement even with disentanglement mechanisms.

Analysis. This failure pattern reveals limitations in:

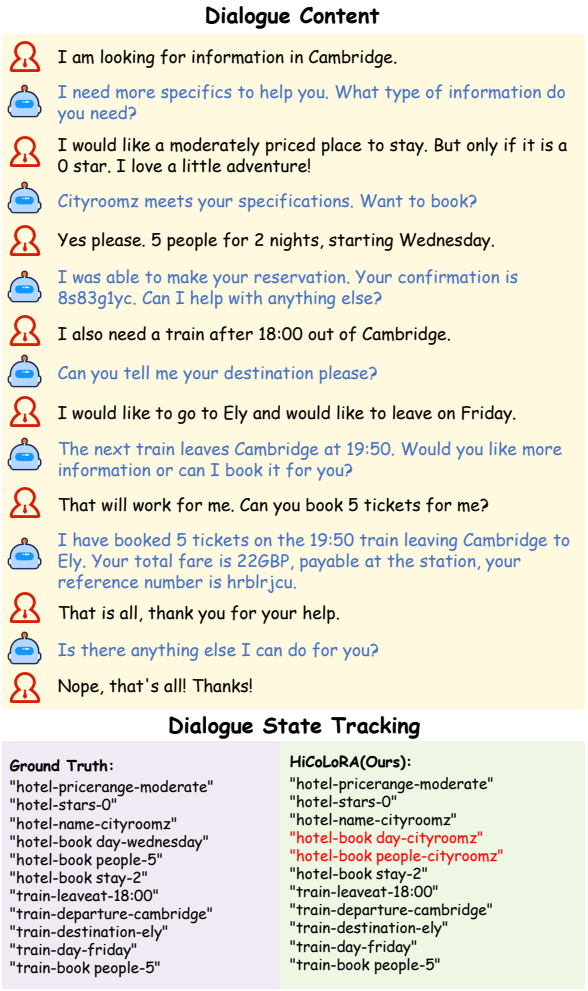


Figure 9: Failure Pattern 2: Cross-Domain Confusion

1. **Domain Disambiguation:** Properly associating slot values with their respective domains in multi-domain dialogs.

2. **Contextual Understanding:** Maintaining clear separation between domain-specific contexts when processing complex multi-domain interactions.

3. **Semantic Overlap Handling:** Dealing with high-overlap domains where lexical similarities between slots from different domains cause confusion. This is particularly challenging when domain-agnostic features are overweighted by the adaptive fusion mechanism.

C.2.3 Pattern 3: Rare Slot Value Cases

Description. These failures occur when the model encounters rare or unseen slot values that were not adequately represented in the training data. Analysis of the MultiWOZ and SGD datasets reveals that such slots are common: in *Attraction*, slots like "entrance fee" and "phone" appear in $<10\%$ of dialogs; in *Hotel*, "stars" and "internet" have fill rates $<20\%$; in *Train*, "trainID" appears in $<5\%$ of dialogs. In a zero-shot setting, HiCoLoRA must generalize to both unseen domains and these rare slot values without any domain specific training examples, presenting a significant challenge.

Example. In MultiWOZ dialogs, HiCoLoRA struggles with predicting rare slot values for specific domains. For instance, in attraction domain dialogs, when users request detailed information about "entrance fee" or "address", the model often fails to correctly predict these values. Similarly, in hotel domain dialogs, when users inquire about specific details like "stars" or "internet", the model shows poor performance. In SGD dialogs, similar patterns emerge. For train domain dialogs, HiCoLoRA often fails to predict "trainID" or "price" information, particularly when these values are not explicitly mentioned in the user utterance but are expected as part of the system response.

Analysis. This failure pattern indicates challenges in:

1. **Rare Value Generalization:** Extending knowledge to handle infrequent slot values that may not have been adequately learned during pre-training. In a zero-shot setting, the model cannot benefit from domain-specific fine-tuning to improve performance on these rare slots.

2. **Contextual Inference:** Properly inferring rare slot values from contextual clues when they are not explicitly mentioned. This is particularly challenging for slots like "trainID" or "reference number" that require the model to generate specific identifiers.

3. **Domain-Aware Initialization:** Current initialization methods (SemSVD-Init) preserve pre-trained knowledge but may not adequately address domain-specific rare slot challenges. Future work could explore domain-aware initialization strategies that better account for rare slot distributions.

4. **Idiosyncratic Semantics Handling:** Dealing with slots that have domain-exclusive terms or idiosyncratic semantics that resist transfer. Spectral clustering may fail for slots with low-frequency terms, and semantic dilution in

higher layers can occur when full collaboration fuses these slots with irrelevant ones.

C.3 Discussion

The case study analysis reveals both the strengths and limitations of HiCoLoRA. The successful cases demonstrate the effectiveness of our hierarchical collaborative architecture, spectral joint clustering, and semantic-enhanced initialization in addressing the core challenges of context-prompt misalignment. However, failure cases highlight areas for future improvement, particularly in handling ambiguous slot boundaries, cross-domain confusion, and rare slot values.

These findings suggest that, while HiCoLoRA represents a significant advance in zs-DST, more research is needed to address the identified failure patterns. Potential directions include:

1. **Enhanced Slot Boundary Detection:** Develop more sophisticated mechanisms to identify and separate slot boundaries in complex utterances.

2. **Improved Domain Disambiguation:** Exploring techniques for better domain separation in multi-domain dialogs.

3. **Rare Value Enhancement:** Investigating data enhancement strategies to improve coverage of rare slot values during training.

In general, the case study provides valuable insight into the practical performance of HiCoLoRA and pinpoints specific future directions, such as developing slot aware refinement techniques to better handle highly idiosyncratic domains.