Reasoning over Object Descriptions Improves Coreference Resolution in Task-Based Dialogue Systems

Oier Ijurco, Oier Lopez de Lacalle
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{oier.ijurco,oier.lopezdelacalle}@ehu.eus

Abstract

Task-based dialogue systems assist users in achieving specific goals, such as executing actions or retrieving information, through natural language interactions. Accurate coreference resolution is essential, as it involves identifying object references within the dialogue—a task that becomes increasingly challenging in visually grounded environments characterized by complex scenes and diverse object metadata. However, coreference resolution in task-based dialogue remains limited by poor generalization across domains and heavy reliance on supervised models that often overfit to dataset-specific artifacts. In this work, we propose a unimodal test-time reasoning approach that enables large language models (LLMs) to reason over detailed object metadata and dialogue history to improve coreference resolution. Empirical results on the SIMMC 2.1 dataset demonstrate that LLMs can generate step-by-step reasoning processes that effectively align dialogue context with objects present in the scene. Extensive experiments highlight the models' ability to link conversations and objects accurately. Moreover, we show that test-time reasoning under few-shot settings generalizes effectively to unseen scenarios and novel objects, outperforming encoder-based supervised methods in cross-domain evaluations. These findings underscore the critical role of structured metadata and careful prompt engineering in enhancing the robustness and generalization of task-oriented dialogue systems.

Keywords: Task-Based Dialogue Systems, Coreference Resolution, Chain-of-Thought, Reasoning

1. Introduction

Task-based dialogue systems aim to assist users in achieving specific goals, such as executing actions or retrieving information, via natural language interaction. In these settings, accurate coreference resolution and entity linking (Ng and Cardie, 2002; Lee et al., 2017) are critical, as systems must correctly identify and resolve references to entities mentioned throughout the dialogue. This challenge is intensified in visually grounded environments, where referenced objects may appear in complex scenes and be described using diverse metadata. Resolving such references requires reasoning over both the dialogue history and multimodal context. As illustrated in Figure 1, effective resolution depends on a comprehensive understanding of the scene, including the description of objects in the metadata.

State-of-the-art coreference resolution methods for task-oriented dialogue systems (Ni et al., 2023) primarily rely on encoder-based architectures fine-tuned on annotated data. While effective within-domain, these models are data-intensive (Huang et al., 2021; Lee et al., 2022) and generalize poorly beyond the training distribution. Recent two-stage approaches, where an encoder processes the dialogue and scene context, followed by a reasoning decoder trained on structured outputs face similar challenges (Long et al., 2023). These systems often overfit to dataset-specific artifacts, such as domain identifiers or object templates, limiting their ability to generalize to unseen objects or domains.

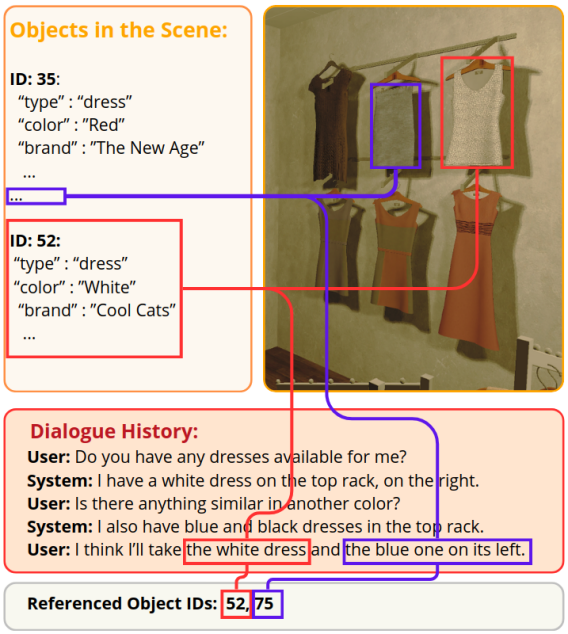

Figure 1: Example of the coreference resolution task in SIMMC 2.1. Given the image of the scene, current object metadata, and dialogue history, the model must identify the object references in the last utterance of the user (e.g. **"the white dress"** with **object id 52**).

As a result, their robustness remains limited, particularly in real-world settings where labeled data is scarce.

To address the challenges of data scarcity and domain dependence, we investigate the use of gen-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2856–2873
11–16 May 2026.

erative large language models (LLMs) for coreference resolution in task-oriented dialogue, grounded in object and scene metadata. Unlike encoder-based architectures, LLMs offer greater flexibility in reasoning over diverse inputs, particularly when structured metadata is combined with natural language prompts. We hypothesize that, when appropriately prompted, LLMs can effectively integrate dialogue context with object and scene descriptions, enabling improved generalization to new domains and scenarios with limited or no annotated data.

Our study focuses on the SIMMC 2.1 dataset (Kottur and Moon, 2023), a benchmark for multimodal, task-oriented dialogue that features diverse scenes, rich object metadata, and multi-turn conversations. Although the dataset is multimodal, we adopt a unimodal (text-only) approach in this work. As illustrated in Figure 1, the coreference resolution task requires the model to identify object references in a user's utterance, given the dialogue history, object metadata, and an image of the scene. We evaluate LLM performance under various prompting strategies (zero-shot, few-shot, and test-time reasoning) and show that test-time reasoning consistently improves accuracy across model variants and settings.

We further investigate cross-domain generalization, finding that fine-tuned LLMs are capable of transferring knowledge across domains. Additionally, ablation studies highlight the importance of different metadata components, and comparisons with state-of-the-art encoder-based models reveal that combining object metadata with test-time reasoning is essential for accurate reference resolution. Notably, LLMs equipped with test-time reasoning exhibit strong performance in scenarios where traditional approaches often fail. As result of our experiments, we find that:

1) Test-time reasoning enhances coreference resolution in task-oriented dialogue: While LLMs often struggle to incorporate rich metadata across tasks, our experiments demonstrate that test-time reasoning significantly improves performance. On the SIMMC 2.1 dataset, LLMs with reasoning prompts produce step-by-step inferences that effectively align dialogue context with object metadata.

2) The proposed approach shows strong cross–domain generalization and transfer capabilities: Cross-domain evaluations reveal that our method generalizes well to unseen objects, addressing key limitations of encoder-based baselines. Additionally, supervised fine-tuning not only enhances in-domain performance but also facilitates cross-domain transfer, with models trained in one domain showing gains in previously unseen domains.

3) Reformulation of structured metadata description into natural language is helpful: We show that the representation of object metadata has a substantial impact on model performance. Specifically, we show that natural language formulations of structured data (e.g. JSON) align better with the model, enabling more effective integration of relational and contextual information.

2. Related Work

Situated Conversational Agents Situated conversational agents interact with users in environments where spatial, visual (Antol et al., 2015; Das et al., 2017), and temporal reasoning (Thomason et al., 2020) is often needed. Recent advances and datasets force models to link entities mentioned in dialogue to those present in a dynamic context. Existing datasets (Das et al., 2017; Saha et al., 2018; Chen et al., 2019; Zhao et al., 2021) challenge models to ground language in multimodal dynamic environments have highlighted the complexity of real-world conversational understanding (Kopp et al., 2005; Alnefaie et al., 2021). However, prior work often uses systems which depend on extensive task-specific supervision and struggle to scale or generalize. To effectively navigate and respond in these complex settings, such agents must understand referring expressions and resolve entities across multiple modalities to maintain coherent conversations. This makes coreference resolution a central capability for such agents.

Coreference Resolution Coreference resolution is a crucial challenge in NLP, with early approaches relying on rule-based or statistical methods (Soon et al., 2001; Ng and Cardie, 2002). With the appearance of neural models, encoder-based architectures have been used to tackle the challenge (Lee et al., 2017). These models perform well on standard textual datasets but struggle in multimodal settings where visual and structured context are critical (Moon et al., 2020). Many multimodal tasks in the literature are related to Multimodal Coreference Resolution, often grouped under the term of Visual Grounding (Sun et al., 2022), such as Referring Expression (Kazemzadeh et al., 2014; Qiao et al., 2020), which involves locating the image region that corresponds to a given textual description. Multimodal Entity Linking further extends the challenges of Coreference Resolution into cross-modal domains, where entities must be linked across text and images (Adjali et al., 2020; Gan et al., 2021; Song et al., 2024; Alonso et al., 2025). Evaluating such complex models and representations requires datasets that reflect the challenges of multimodal reasoning and entity linking.

Task-oriented Dialogue Datasets Datasets like MultiWOZ (Eric et al., 2019), SIMMC (Kottur and Moon, 2023), and challenges like DSTC (Hender-

son et al., 2014; Williams et al., 2014, 2016) have driven research in task-oriented dialogue by providing annotated multi-turn conversations. SIMMC 2.1 in particular introduces a multimodal setup, making it a rich benchmark for evaluating multimodal coreference. These datasets highlight the limitations of static, fully trained models and expose the need for dynamic reasoning capabilities, particularly when facing out-of-distribution examples at inference. This motivates recent interest in context engineering strategies that adapt models to novel situations on the fly.

Context Engineering Recent work highlights context engineering (Mei et al., 2025) as a key paradigm for improving model robustness and adaptability without retraining. Techniques like test-time training (Sun et al., 2020; Akyürek et al., 2024), retrieval-augmented generation (Lewis et al., 2020; Gao et al., 2023), and in-context learning (Brown et al., 2020; Dong et al., 2024) use available inputs to adapt models during inference. In dialogue systems, such mechanisms are under-explored. However, they hold significant promise in enabling conversational agents to better resolve references, reason about new contexts, and personalize responses, all of which are vital for situated and multimodal interactions.

3. Problem Formulation

SIMMC 2.1 Dataset We base our study on the SIMMC 2.1 dataset (Kottur et al., 2021; Kottur and Moon, 2023), a task-oriented dialogue corpus designed for multimodal assistant-user interactions in realistic shopping scenarios. The dataset features complex scenes with an average of 19.7 objects arranged realistically, making it an ideal testbed for robust coreference resolution.

The dataset comprises two distinct domains: Fashion and Furniture. While both domains share a common structure, each introduces unique objects and scenarios that do not appear in the other domain. This domain-specific variation enables a controlled evaluation of model generalization, requiring systems to adapt to novel object types and contextual cues.

Each dialogue in the dataset is grounded in a visual environment and includes a combination of images, 3D spatial coordinates, and detailed object metadata. Dialogues are multi-turn and task-driven, encompassing user intents such as item requests, comparisons, and attribute-based inquiries. In total, the dataset comprises 11,244 dialogues and approximately 117,000 utterances.

Formally, a dialogue can be represented as a sequence of tuples: $D := (U_t, A_t, M_t, S_t, B_t)_{t=1}^{r}$, where U_t and A_t denote the user and assistant utterances at turn t. M_t represents the multimodal context, including indices of referred objects, and S_t encodes the scene context, such as object attributes and spatial metadata (see §4 for details). The belief state B_t includes dialogue acts and slot-value pairs, but is not utilized in this work.

Although the benchmark defines four tasks, in this work, we focus on the coreference resolution task, which is detailed in the following section.

Coreference Resolution Task In SIMMC, the task requires identifying which objects, if any, are referenced in the user's most recent utterance within a multi-turn dialogue. Each dialogue is grounded in a scenario with a scene S, containing candidate objects $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$, each associated with structured metadata meta(o_i), a dialogue history $H = \{u_1, s_1, \ldots, u_{t-1}, s_{t-1}\}$, where u_i and s_i are the user and system utterances at turn i, and the current user utterance u_t. The task is to predict, for each object $o_i \in \mathcal{O}$, a label $y_i \in \{0, 1\}$ indicating whether o_i is referred to in u_t given H, S and meta(o_i). If no object is referenced, then $y_i = 0$ for all i. This frames the coreference resolution task as a per-object binary classification problem within each scene. Accurate resolution requires joint reasoning over the dialogue context, multimodal scene information, and fine-grained object metadata.

4. Proposed Approach

We adopt a generative large language model (LLM) based approach that reframes the problem of coreference resolution problem as a reference identification task. Contrary to the previous approaches, in which they focus on fine-tuning models on annotated data, we adopt in-context learning and chain-of-thought approaches to fully harness the reasoning capabilities of current LLMs. Our approach allows us to generalize the coreference resolution task to unseen objects and scenes, which is exemplified in Figure 2, and proceeds as follows. At each dialogue turn, the LLM is prompted with information that includes: 1) A complete list of objects in the scene, each with detailed **metadata** such as the `ID`, `type`, `color`, `brand`, and `location`, among others. 2) The **dialogue history** between the user and the system. The dialogue history comes with all the object IDs mentioned earlier in the conversation[1]. Each turn is finished with the last utterance of the user. 3) Finally, **task-specific instructions** outlining the required actions for the model to complete the task are given. In addition, a structured set of reasoning steps is provided to aid the model in aligning contextual information appearing in the dialogue and the detailed description of objects occurring in the scene. Further details of the reason-

[1] Previously mentioned objects are coded as `<SOM> objectID <EOM>`.

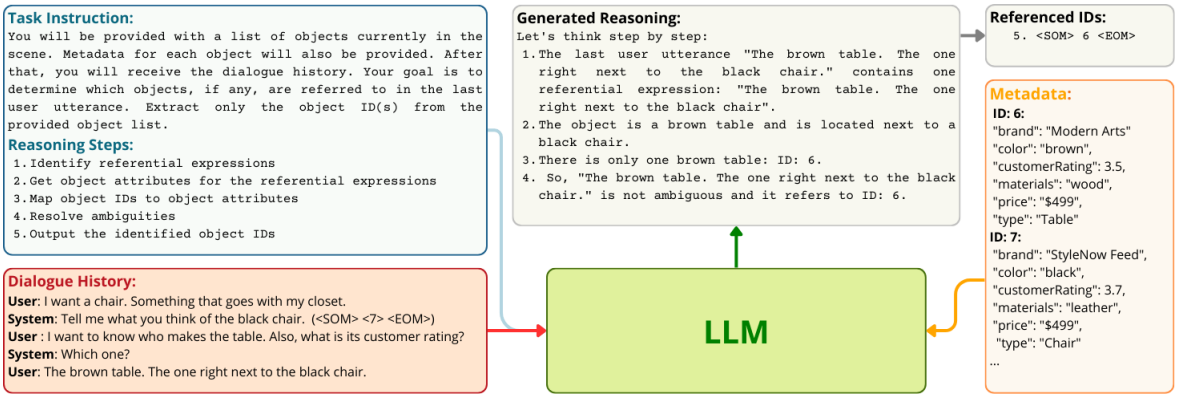

Figure 2: Example of our approach for coreference resolution task in SIMMC 2.1. The LLM receives as input instruction to solve the task (**Task Instruction**) along with some reasoning steps that incorporate object descriptions (**Metadata**) and the previous dialogue and object references (**Dialogue History**). The LLM generates some reasoning that delivers the referenced object IDs by the user (**Referenced IDs**). In this case, the user is asking for information about *the brown table* that is linked to `ID:6` in metadata. The LLM should return `<SOM>6<EOM>` after reasoning about the input.

ing steps are given in the follow-up section (§4.1).

The LLM is then prompted to output the object ID(s) mentioned in the last user utterance. In case the user does not mention any object, the model is instructed to output an empty sets of objects (i.e., `<SOM><EOM>`)[2]. The response is also considered as type of prediction over the set of possible object IDs.

The proposed formulation allows the model to reason directly over the metadata and conversational context without any fine-tuning on annotated data. We experimented with various prompting strategies, including few-shot examples and chain-of-thought reasoning to enhance the performance of the model. The approach is flexible, domain-agnostic, and well-suited to scenarios where traditional models struggle due to limited supervision or unseen objects in a new domain.

4.1. Reasoning steps

Alignment of the conversational context and information contained in the metadata is not trivial for the current LLMs, but it is required to correctly reference the mentioned objects. Thus, the aim of this additional prompt is to generate a concrete inferring process step-by-step and implement corresponding actions on the conversational history and metadata to get information about object IDs. As Figure 2 shows, the LLM takes the task instructions, the reasoning steps, and generates the reasoning that aligns dialogue history and metadata to provide the referenced object ID. The LLM needs to go through the subsequent step-by-step deliberation process of 5 steps:

1. **Identification of referential expressions**: In this step, the model extracts the noun phrases, pronouns and any descriptive phrases in the dialogue that may mention any object of the scene.

2. **Getting object attributes for the referential expressions**: In this step, the goal is to gather relevant metadata (e.g., type, color, location) for each identified object mention.

3. **Mapping the current object IDs to object attributes**: Once the relevant object attributes are obtained, the model attempts to match object IDs to the extracted attributes, selecting only those with strong alignment.

4. **Ambiguity resolution**: In cases where multiple objects meet the referential criteria, the model incorporates dialogue history to resolve ambiguity and narrow down the selection.

5. **Output the identified object IDs**: Finally, the model returns confident object IDs between `<SOM>` and `<EOM>` tags. Otherwise, it outputs nothing to avoid errors.

These reasoning steps reflect the way a human would reason to perform this same task and predict the objects that are being referenced in the given utterance in the dialog.

5. Experimental Setting

We conducted our experiments using a variety of strong open models that are publicly available. The evaluated models include the Llama 3 family (Grattafiori et al., 2024), which we tested Llama

[2]E.g., "`User:` *Hi, do you have any jackets today?*" would be an utterance without mentioning any object, in which empty set of object IDs should be returned.

3.1-8B and 3.3-70B models, instructed Gemma-7B (Team et al., 2024), instructed Mistral-7B (Jiang et al., 2023), instructed Qwen 2.5-7B (Qwen et al., 2025), Qwen3-4B (Yang et al., 2025) and the distilled version of DeepSeek-R1 to Llama3.1-8B (DeepSeek-AI et al., 2025).

All open-weight models were evaluated in a computing cluster using four NVIDIA A100 80GB GPUs. The entire set of experiments was estimated to have required approximately 750 GPU hours.

We conducted two main sets of experiments to evaluate the effectiveness of the large language models for coreference resolution in task-oriented dialogue settings. The first set of experiments is devoted to assessing the contribution of test-time reasoning to understanding object descriptions in the metadata and the dialogue history (§6.1). We explore different prompting strategies to measure the effectiveness of test-time reasoning. In addition, to ensure the models leverage object descriptions effectively, we conduct an ablation study on the availability of information sources, such as previous object references in the dialogue and access to the object's metadata (§7.1).

In the second set of experiments, we measure the ability of the LLMs to generalize across domains and handle unseen objects (§6.2). To this end, we perform cross-domain evaluations, in which models are trained in one domain (say furniture domain) and evaluated in the other domain (say fashion). This setting ensures no seen objects occur in the testing domain. In addition, we compare the generative models against state-of-the-art encoder models under the same conditions to measure their robustness and adaptability to new, unseen domains. All experiments were performed using the SIMMC 2.1 dataset, focusing on both the Fashion and Furniture domains.

To efficiently fine-tune the model in the second set of experiments, we employ parameter-efficient fine-tuning (PEFT) (Han et al., 2024) using LoRA (Low-Rank Adaptation) (Hu et al., 2022). Specifically, we fine-tune the Qwen2.5-7B-Instruct and the Llama3.3-70B-Instruct models with LoRA configurations: rank (r) = 4, LoRA alpha = 8, and dropout = 0.05.

The dataset includes approximately 11K task-oriented dialogues (around 117K utterances) between an assistant and a user, situated in commercial store environments enriched with scene images and item metadata. The dataset spans the fashion (7.2K dialogues) and furniture (4K dialogues) domains, with the fashion domain presenting more complexity due to higher item density and visual overlap in scenes. On average, dialogues contain 10.4 utterances, with assistant turns averaging 13.7 words. Each dialogue mentions about 4.7 objects, while scenes contain an average of 19.7 objects. The average number of objects per furniture scene is slightly over 10, whereas in the fashion domain the mean exceeds 30, reaching over 55 items in some cases. The data is split into Train (64%, 7307 dialogues), Dev (5%, 563), Dev-test (15%, 1687), and Std-test (15%, 1287). We use the Dev-test set for all our evaluations, which comprises 8609 instances.

In all the experiments, the few-shot examples have been selected from the training split to reflect a diverse range of scenarios. This ensures that the model is exposed to various challenges it may encounter during inference. Specifically, the examples include: one instance where no objects are referenced, another where the currently referred objects have been previously mentioned in the dialogue, and a final case where the current objects are being referred to for the first time. The examples span across both the Fashion and Furniture domains, further contributing to a more representative and balanced demonstration of the model's capabilities.

Additionally, we conduct a qualitative analysis to better understand the factors influencing model behavior (§7.1). Specifically, we examine how input types, metadata formats, and prompting strategies affect performance and interpretability. These analyses include evaluating information access in resolving references, comparing metadata formats and assessing prompt design effects on predicted object references. Together, these studies offer deeper insights into the interpretability of LLM behavior in task-oriented dialogue systems.

6. Results

6.1. The Impact of Reasoning

In the subsequent experiments, we evaluate the effectiveness of test-time reasoning by comparing our approach against different prompt completion strategies. We consider three types of prompts: **Zero-shot**, where the LLM receives only the task instructions along with dialogue history and object metadata, but no input-output examples; **Few-shot**, which extends the zero-shot setting by including three coreference resolution examples from both the fashion and furniture domains; and **Few-shot+Reasoning**, in which these few-shot examples are augmented with an additional reasoning step that explicitly explains how object metadata relates to the dialogue context. The full prompt structure and a complete few-shot instance used in our test-time reasoning approach are provided in Appendix A.

Table 1 reports the precision, recall, and F1 scores obtained in SIMMC 2.1 for several models evaluated across the different prompt settings.

	Model	Size	Precision	Recall	F1 Score
	Random		2.68	49.87	5.09
Zero-shot	Qwen3	4B	**29.01**	63.91	39.90
	Gemma	7B	16.93	52.74	25.64
	Mistral	7B	25.70	66.62	37.09
	Qwen2.5	7B	24.14	71.66	36.12
	Llama3.1	8B	27.74	63.03	38.53
	DS-R1	8B	22.98	66.77	34.19
	Llama3.3	70B	26.61	**80.10**	**39.95**
Few-shot	Qwen3	4B	39.13(+10.12)	59.15(-4.76)	47.10(+7.2)
	Gemma	7B	19.58(+2.65)	41.38(-11.36)	26.58(+0.94)
	Mistral	7B	29.14(+3.44)	62.65(-3.97)	39.77(+2.68)
	Qwen2.5	7B	28.84(+4.7)	70.05(-1.61)	40.86(+4.74)
	Llama3.1	8B	30.34(+2.6)	68.39(+5.36)	42.03(+3.5)
	DS-R1	8B	25.31(+2.33)	67.17(+0.4)	36.77(+2.58)
	Llama3.3	70B	**39.32**(+12.71)	**73.12**(-6.98)	**51.14**(+14.34)
Reasoning	Qwen3	4B	**59.98**(+20.85)	45.21(-13.94)	51.55(+4.45)
	Gemma	7B	37.86(+18.28)	43.24(+1.86)	40.37(+13.79)
	Mistral	7B	39.82(+10.68)	52.93(-9.72)	45.45(+5.68)
	Qwen2.5	7B	49.90(+21.06)	57.32(-12.73)	53.35(+12.49)
	Llama3.1	8B	52.16(+21.82)	59.23(-9.16)	55.47(+13.44)
	DS-R1	8B	53.50(+28.19)	43.92(-23.25)	48.24(+11.47)
	Llama3.3	70B	58.73(+19.41)	**61.37**(-11.75)	**60.02**(+8.88)

Table 1: Results of the models evaluated in different prompt settings (zero-shot, few-shot, and test-time reasoning). Numbers in parentheses show the differences with the previous prompt technique: zero-shot vs few-shot, and few-shot vs reasoning. DS-R1 refers to DeepSeek-R1-Distill-Llama-8B

Few-shot improvement In terms of the F1 score, the results demonstrate a consistent improvement when examples are incorporated into the prompt. Note that only 3 examples are added in the prompt. Overall, all models with 7B to 8B parameters exhibit an improvement between 1 and 4.5 points compared to zero-shot versions. The 70B Llama model shows an even greater enhancement, achieving an increase of more than 14 points. The performance of Qwen3 is particularly noteworthy: despite being significantly smaller in size, it achieves a comparable F1 score to the 70B Llama model in the zero-shot setting. Moreover, Qwen3 exhibits a significant improvement of 7.2 points when few-shot examples are incorporated into the prompt. Analysis indicates that the improvement in F1 is primarily driven by an increase in precision, at the cost of recall. This suggests that, under the few-shot setting, the models tend to respond to fewer IDs and have begun to better infer how to perform the task.

Few-shot vs Reasoning The application of reasoning on top of few-shot learning also yields positive results in terms of F1 scores. All evaluated models consistently demonstrate significant improvements, with the observed gains being even more pronounced than the differences between zero-shot and few-shot settings. Across all models, the average improvement is approximately 10 points. Notably, the Qwen2.5 and Llama3.1 models exhibit particularly strong improvements, achieving performance levels close to that of the 70B Llama3.3 model. In terms of precision, the improvements are substantial, with the models achieving an average increase of 19 points. However, this precision gain comes at the cost of Recall, as reasoning trends to reduce overprediction of objects (§7.3).

Model size Although a detailed analysis has not been conducted, the experiments suggest that model size plays an important role. As the model size increases, the ability to correctly perform reasoning steps also improves. Despite being specifically designed for reasoning, Deepseek-R1 does not achieve strong results.

6.2. Cross-domain Experiments

The following experiments investigate the cross-domain generalization capabilities of LLMs, with a particular emphasis on their performance when exposed to novel scenes and entities absent during training. Specifically, we conduct two evaluations: in the first, the model is trained solely on data from the fashion domain and tested on the furniture domain (**Fashion** $\rightarrow$ **Furniture**); in the second, the training and testing domains are reversed, with the model trained on furniture data and evaluated on fashion (**Furniture** $\rightarrow$ **Fashion**). In both cases, the test sets contain previously unseen objects, enabling a robust assessment of the model's generalization across domains. We selected Qwen2.5 and Llama3.3 models for our experiments. We compare the generative models against a strong encoder-based model ($\text{BART}_{\text{encoder}}$) under the same conditions to measure robustness. The results of the experiments are shown in Table 2a and Table 2b.

Tables are divided into prompt variants of few-shot (FS) and supervised fine-tuning (SFT). In both cases, reasoning is applied at inference time too (results of models with subscript+R). For fine-tuning, models are provided with task instructions, metadata, and dialogue history, and they learn to generate mentioned object identifiers.

Overall, both domains exhibit similar behaviors of models. Due to the fewer objects to predict in the Furniture domain, all models tend to achieve higher performance in this context. The results indicate that supervised classifiers based on encoder architectures ($\text{BART}_{\text{encoder}}$) cannot generalize effectively to out-of-domain scenarios. While these models obtain strong results in in-domain evaluations (achieving F1 scores of approximately 71 and 79 in the Fashion and Furniture domains, respectively), their performance degrades substantially when applied to unseen domains. In contrast, models based on few-shot learning and reasoning (as shown in the FS rows) demonstrate stronger cross-domain robustness.

The few-shot results reveal a clear imbalance between the two transfer directions. When models

(a) Trained on Fashion			
Setting	**Model Name**	**Fashion**	**Furniture**
	$BART_{encoder}$	70.93	6.41
FS	Qwen2.5	39.86	49.85
	Llama3.3	52.05	55.54
	$Qwen2.5_{+R}$	52.60	63.22
	$Llama3.3_{+R}$	59.96	59.28
SFT	Qwen2.5	55.98	67.06
	Llama3.3	**73.62**	**75.51**
	$Qwen2.5_{+R}$	55.04	64.92
	$Llama3.3_{+R}$	66.06	65.04
(b) Trained on Furniture			
Setting	**Model Name**	**Fashion**	**Furniture**
	$BART_{encoder}$	5.44	**78.78**
FS	Qwen2.5	35.27	52.63
	Llama3.3	46.46	56.09
	$Qwen2.5_{+R}$	45.18	60.59
	$Llama3.3_{+R}$	54.80	60.85
SFT	Qwen2.5	51.34	63.05
	Llama3.3	59.66	64.60
	$Qwen2.5_{+R}$	49.38	60.47
	$Llama3.3_{+R}$	**62.55**	60.44

Table 2: Results of models trained on (a) the **Fashion** domain and (b) the **Furniture** domain. Each is evaluated both in-domain and cross-domain. **FS** refers to the few-shot setting (no actual training). **SFT** refers to supervised fine-tuning. The subscript **+R** indicates test-time reasoning is applied.

are given examples from the Fashion domain, they achieve higher scores when evaluated on Fashion than the opposite direction. This suggests that the Fashion domain, being more complex, provides richer examples. As a result, models exposed to Fashion few-shot examples learn more generalizable patterns that transfer effectively to the simpler Furniture domain. On the other hand, few-shot learning on Furniture lacks sufficient representational variety to generalize back to Fashion, leading to lower cross-domain scores. These findings highlight that domain complexity plays an important role in enabling LLMs to build transferable understanding from few-shot examples.

Furthermore, fine-tuned LLMs (SFT-Qwen and SFT-Llama) yield significant improvements in both in-domain and cross-domain performance over their few-shot counterparts. The Llama models, in particular, achieve the best overall results, reaching F1 scores above 73 and 75 when trained on Fashion and tested on Fashion and Furniture, respectively (Table 2a). Improvements are also observed when training on Furniture and testing on Fashion, where SFT-Llama rises from 49.78 to 59.66. Similar trends are observed for Qwen models, though with slightly lower absolute performance. It is worth noting that inference-time reasoning does not yield additional benefits for fine-tuned models, and in some cases slightly decreases performance (SFT-Llama+R and SFT-Qwen+R).

7. Analysis

7.1. Effect of Information Access

We examine the role of object descriptions in enhancing the ability of LLMs to resolve referential expressions. To assess the contribution of different information sources, we conduct a controlled ablation study across the different prompt settings. Specifically, we evaluate model performance under three input configurations: in the **All information** setting, the model receives full access to both the descriptions of all objects in the scene and the complete dialogue history; in the **No Metadata** setting, object metadata is withheld, so the model relies solely on the dialogue history to resolve references; and in the **No Object References** setting, mentions of objects within the dialogue history are removed, limiting the model's input to metadata and non-referential dialogue content. Example prompts for each configuration are shown in Appendix B.

Setting	**Metadata**	**Precision**	**Recall**	**F1 Score**
Zero-shot	All Info	24.14	71.66	36.12
	No Metadata	22.18	52.96	31.27
	No Objects	16.75	60.82	26.26
Few-shot	All Info.	28.84	70.54	40.86
	No Metadata	36.56	51.98	42.93
	No Objects	20.66	56.93	30.31
FS+Reasoning	All Info	49.90	57.32	53.35

Table 3: Qwen2.5-7B performance across different metadata representations and settings.

Setting	**Metadata**	**Precision**	**Recall**	**F1 Score**
Zero-shot	All Info	26.61	80.10	39.95
	No Metadata	27.94	55.35	37.13
	No Objects	23.64	72.49	35.66
Few-shot	All Info	39.32	73.12	51.14
	No Metadata	56.87	45.43	50.51
	No Objects	34.69	70.70	46.54
FS+Reasoning	All Info	58.73	61.37	60.02

Table 4: Llama3.3-70B performance across different metadata representations and settings.

Our ablation analysis underscores the critical role of both object metadata and previously mentioned objects within the dialogue. The results in Tables 3 and 4 suggest that LLMs mainly rely on previously mentioned objects to resolve references effectively. In the zero-shot setting, removing these references reduces Qwen's F1 score by about 10 points, while the larger Llama 70B model, although more robust, still shows a notable drop. This indicates that even larger models remain sensitive to referential context.

A comparable pattern appears in the few-shot setting: removing object references lowers Qwen's score by around 11 points and Llama's by about 5, confirming that reference tracking strongly influences performance regardless of model scale or supervision.

When object metadata is excluded, performance differences narrow, suggesting that models may not effectively integrate structured metadata with dialogue context. For instance, in the few-shot setting, Llama's performance remains stable and Qwen's slightly improves, implying that metadata can sometimes introduce noise without explicit prompt guidance.

7.2. Metadata Representation

An insight from our experiments is that the way object metadata is represented substantially affects model performance. As shown in Figure 3, expressing metadata in natural language outperforms structured formats such as JSON. While JSON provides a clear schema, it can hinder cross-attribute reasoning due to tokenization inefficiencies and limited contextual integration.

In contrast, natural language formulations align better with the model's pretraining, enabling more effective integration of relational and contextual information. We compare three variants of metadata representations: in the first, **Raw coordinates** are encoded as normalized numerical values, an exact format that, however, forces the model to infer spatial semantics from raw numbers. The second representation, **Coordinates as natural language**, converts numerical values into coarse spatial descriptors such as "bottom-left" or "center", introducing semantically meaningful spatial cues about object location. The third variant, **Full natural language descriptions**, reformulates all metadata fields into fluent, human-readable sentences, removing key-value structures and providing the model with a more natural input format. Concrete examples of all three representation types are given in Appendix C.

As illustrated in Figure 3, the full natural language representation consistently outperforms the other formats, suggesting that LLMs are more adept at reasoning over unstructured, linguistically rich inputs than over structured data.

7.3. Number of Returned Objects

Figure 4 presents the impact of different prompting strategies on the number of predicted object references. We observe that prompts incorporating few-shot examples and explicit reasoning steps lead to fewer object IDs being predicted per utterance. This reduction is associated with increased precision and a slight decrease in recall, suggesting that

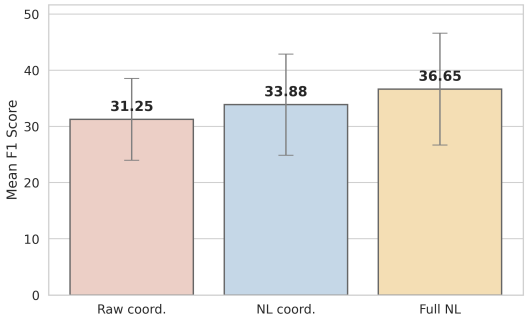

Figure 3: F1 mean scores and standard deviation of all models across different object representation types. Raw coordinates use normalized values for coordinates; NL coordinates map these values to spatial descriptors; Full NL expresses all metadata in fluent natural language.

the model becomes more conservative in its predictions. Rather than over-generating references, the model appears to better recognize uncertainty and abstains from producing potentially incorrect outputs. This behavior indicates a stronger alignment with task semantics, favoring precision over recall in ambiguous contexts.

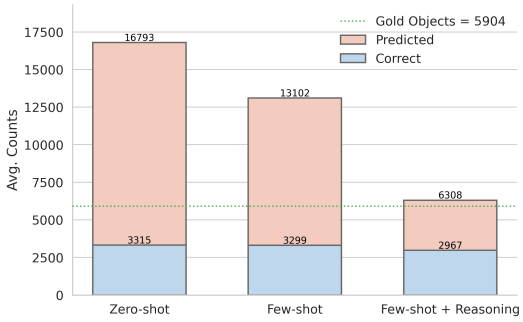

Figure 4: The average count of predicted objects and correct predictions per setting. The average is computed across all the models and settings used in the experiments.

8. Conclusions

We present a prompting-based approach for resolving object references in task-oriented dialogue by leveraging rich object metadata, structured scene context, and carefully designed prompting strategies. Our experiments demonstrate that large language models can generate effective step-by-step reasoning to align dialogue history with objects in the environment. We show that test-time reasoning combined with few-shot examples significantly improves coreference resolution performance. Cross-domain evaluations further show that our method generalizes well to unseen objects, addressing key limitations of encoder-based baselines. Addi-

tionally, supervised fine-tuning not only enhances in-domain performance but also facilitates cross-domain transfer, with models trained in one domain showing gains in previously unseen domains. Our analysis highlights the importance of representing object metadata in natural language, which better aligns with model pretraining and supports more effective integration of relational and contextual cues.

Future work may explore fine-tuning strategies that explicitly target reasoning capabilities, as well as methods for prompt compression and dynamic context selection. Integrating visual grounding modules also presents a promising way to improve performance in interactive, multimodal settings.

9. Limitations

While our approach shows substantial improvements in coreference resolution by leveraging large language models and structured metadata, several limitations remain, highlighting opportunities for future work.

First, our experiments are conducted exclusively on the SIMMC 2.1 dataset. Although SIMMC 2.1 provides a rich multimodal setting, the lack of comparable publicly available datasets restricts the opportunity to evaluate our findings to broader task-oriented dialogue scenarios.

Second, despite SIMMC 2.1's multimodal nature, our method focuses solely on textual and structured metadata, omitting direct use of visual input. Incorporating visual signals into the reasoning process remains an important direction for future research.

Finally, while our results with supervised fine-tuning suggest promising cross-domain transfer capabilities, further investigation is needed to better understand the underlying mechanisms and to improve fine-tuning strategies for robustness and scalability.

Acknowledgments

This work was supported by the European Research Council under Horizon Europe, grant number 10113572, related to LUMINOUS project, and the Spanish Ministry of Science and Innovation (AI4I/MOLVI project PID2024-157855OB-C32 and HumanAIze project AIA2025-163322-C61) funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU.

10. Bibliographical References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *European Conference on Information Retrieval*, pages 463–478. Springer.

Ekin Akyürek, Mehul Damani, Adam Zweiger, Linlu Qiu, Han Guo, Jyothish Pari, Yoon Kim, and Jacob Andreas. 2024. The surprising effectiveness of test-time training for few-shot learning. *arXiv preprint arXiv:2411.07279*.

Ahlam Alnefaie, Sonika Singh, Baki Kocaballi, and Mukesh Prasad. 2021. An overview of conversational agent: applications, challenges and future directions. In *17th International Conference on Web Information Systems and Technologies*. SCITEPRESS-Science and Technology Publications.

Iñigo Alonso, Ander Salaberria, Gorka Azkune, Jeremy Barnes, and Oier Lopez de Lacalle. 2025. Vision-language models struggle to align entities across modalities. *arXiv e-prints*, pages arXiv–2503.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2019. The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service. *arXiv preprint arXiv:1911.09969*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

DeepSeek-AI, Daya Guo..., and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal entity linking: a new dataset and a baseline. In *Proceedings of the 29th ACM international conference on multimedia*, pages 993–1001.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri..., and Zhiyu Ma. 2024. The llama 3 herd of models.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yichen Huang, Yuchen Wang, and Yik-Cheung Tam. 2021. Uniter-based situated coreference resolution with rich multimodal input. *arXiv preprint arXiv:2112.03521*.

Albert Q. Jiang, Alexandre Sablayrolles..., and William El Sayed. 2023. Mistral 7b.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Stefan Kopp, Lars Gesellensetter, Nicole C Krämer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide–design and evaluation of a real-world application. In *Intelligent Virtual Agents: 5th International Working Conference, IVA 2005, Kos, Greece, September 12-14, 2005. Proceedings 5*, pages 329–343. Springer.

Satwik Kottur and Seungwhan Moon. 2023. Overview of situated and interactive multimodal conversations (simmc) 2.1 track at dstc 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 235–241.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Jinhyeon Kim, Youngjune Lee, Ran Han, Yoonhyung Kim, Minho Park, Kangwook Lee, Haebin Shin, et al. 2022. Tackling situated multi-modal task-oriented dialogs with a single transformer model. In *Proc. AAAI Conf. Artif. Intell. Workshop*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yuxing Long, Huibin Zhang, Binyuan Hui, Zhenglu Yang, Caixia Yuan, Xiaojie Wang, Fei Huang, and Yongbin Li. 2023. Improving situated conversational agents with step-by-step multi-modal logic reasoning. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 15–24.

Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. 2025. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*.

Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, et al. 2020. Situated and interactive multimodal conversations. *arXiv preprint arXiv:2006.01460*.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 104–111.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A

systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440.

Qwen, An Yang..., and Zihan Qiu. 2025. Qwen2.5 technical report.

Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Shezheng Song, Shan Zhao, Chengyu Wang, Tianwei Yan, Shasha Li, Xiaoguang Mao, and Meng Wang. 2024. A dual-way enhanced framework from text matching point of view for multimodal entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19008–19016.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR.

Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. 2022. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International conference on Multimedia*, pages 404–412.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR.

Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The dialog state tracking challenge series. *AI Magazine*, 35(4):121–124.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Nan Zhao, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. The jddc 2.0 corpus: A large-scale multimodal multi-turn chinese dialogue dataset for e-commerce customer service. *arXiv preprint arXiv:2109.12913*.

A. Prompt used for our test-time reasoning approach

Figure 5 shows the prompt used in our test-time reasoning approach, which is structured to guide a language model through a multi-step inference process aimed at grounding user utterances in object metadata. It is divided into clearly demarcated sections: Instructions, reasoning steps, and few-shot examples. The instructions define the task, constraints, and expected output format. The reasoning steps section outlines a step-by-step procedure designed to encourage chain-of-thought reasoning, including referential expression identification, attribute mapping, and ambiguity resolution. An example of a few-shot instance can be seen in Figure 6.

B. Ablation Study: Input Configurations

This section presents an example for an instance shown under each of the different experimental setups used in our ablation study to assess how varying levels of contextual information impact object reference resolution. Each setting manipulates the type and amount of information available to the model, allowing us to isolate the contribution of different informational cues. The full configuration retaining all metadata and object references is shown in Figure 7, while Figure 8 and Figure 9 illustrate the effect of omitting object metadata and object references, respectively.

C. Object Representation Types

This section shows an example of how the same object is represented using the three different ways of representing an object that we have used in our experiments: structured with raw coordinates (Figure 10), structured with natural language location information (Figure 11), and fully naturalized as descriptive text (Figure 12).

```
Your task is to identify object IDs referenced in the last user utterance.

### Instructions:
- You will be provided with a list of objects currently in the scene. Metadata for each object
    will also be provided.
- After that, you will receive the dialogue history.
- Your goal is to determine which objects, if any, are referred to in the last user utterance.
- Extract only the **object ID(s)** from the provided object list.
- **By default, assume no object is referenced** unless you are **highly confident**.
- Output object ID(s) **only when absolutely certain** that the last user utterance refers to
   them.
- If any objects are referenced in the last user utterance, output their ID(s) **between `<SOM
   >` and `<EOM>`**.
- **Use only the provided object IDs and metadata.** If an object is not listed, ignore it.
- Whenever an object ID appears inside `<SOM> ... <EOM>` in the dialogue history, it **
   explicitly marks the objects mentioned in that part of the conversation.**
- **Do NOT generate explanations, apologies, or extra text.**
- **Do NOT answer any questions in the conversation.**
- Format: `<SOM> ID1, ID2 <EOM>` (Example: `<SOM> 55, 85 <EOM>`).
- **If unsure or if the last user utterance is ambiguous, do NOT output any object IDs.**

### Reasoning Steps:
1. **Identify Referential Expressions:** Extract noun phrases, pronouns, or descriptive
   phrases that may refer to objects.
2. **Get Object Attributes for Referential Expressions:** Get and list the metadata from these
    expressions (e.g., type, color, location).
3. **Map Current Object IDs to Object Attributes:** Only select objects when there is strong
   alignment between the referential expressions and object metadata.
4. **Resolve Ambiguity:** If multiple objects match a description, consider context from
   dialogue history.
5. **Output Identified Object IDs:** If confident, format IDs between `<SOM>` and `<EOM>`,
   otherwise output nothing.

{Few-shot examples, their object metadata, dialogue history and expected reasoning steps}

### Current Objects in Scene:
{List of Current Objects in Scene}

### Dialogue History:
{Current Dialogue History}

#### Reasoning Steps:
Let's think step by step:
```

Figure 5: Full prompt used for our test-time reasoning approach. The prompt is structured into three sections: task instructions and output format constraints, a chain-of-thought reasoning procedure, and few-shot examples.

```
#### Current Objects in Scene:
ID: 0 : {"description": "white EndTable from the brand StyleNow Feed. Made from wood. It has a
    customer rating of 4.9 out of 5. Priced at $399.00.", "location": "left"}
ID: 1 : {"description": "white EndTable from the brand StyleNow Feed. Made from wood. It has a
    customer rating of 4.9 out of 5. Priced at $399.00.", "location": "left"}
ID: 2 : {"description": "wooden Table from the brand River Chateau. Made from wood. It has a
    customer rating of 4.3 out of 5. Priced at $399.00.", "location": "left"}
ID: 3 : {"description": "brown Chair from the brand Modern Arts. Made from leather. It has a
    customer rating of 4.2 out of 5. Priced at $299.00.", "location": "left"}
ID: 4 : {"description": "brown CouchChair from the brand Downtown Stylists. Made from leather.
    It has a customer rating of 2.9 out of 5. Priced at $349.00.", "location": "left"}
ID: 5 : {"description": "grey AreaRug from the brand North Lodge. Made from wool. It has a
    customer rating of 3.8 out of 5. Priced at $499.00.", "location": "top-left"}
ID: 6 : {"description": "brown Table from the brand Modern Arts. Made from wood. It has a
    customer rating of 3.5 out of 5. Priced at $499.00.", "location": "left"}
ID: 7 : {"description": "black Chair from the brand StyleNow Feed. Made from leather. It has a
    customer rating of 3.7 out of 5. Priced at $499.00.", "location": "center"}
ID: 8 : {"description": "blue AreaRug from the brand River Chateau. Made from wool. It has a
    customer rating of 3.4 out of 5. Priced at $249.00.", "location": "center"}
ID: 9 : {"description": "blue AreaRug from the brand River Chateau. Made from wool. It has a
    customer rating of 3.4 out of 5. Priced at $249.00.", "location": "left"}
ID: 10 : {"description": "red AreaRug from the brand Art Den. Made from natural fibers. It has
    a customer rating of 3.8 out of 5. Priced at $199.00.", "location": "center"}

#### Dialogue History:
User : I want a chair. Something that goes with my closet. System : Tell me what you think of
    the black chair. <SOM> <7> <EOM> User : I want to know who makes the table. Also, what is
    its customer rating? System : Which one? <SOM>  <EOM> User : The brown table. The one
    right next to the black chair.

#### Reasoning Steps:
Let's think step by step:
1. The last user utterance "The brown table. The one right next to the black chair." contains
    one referential expression: "The brown table. The one right next to the black chair".
2. The object is a brown table and is located next to a black chair.
3. There is only one brown table: ID: 6.
4. So, "The brown table. The one right next to the black chair." is not ambiguous and it
    refers to ID: 6.
5. <SOM> 6 <EOM>
```

Figure 6: Illustration of a complete few-shot instance as provided in the prompt, including the user utterance, object metadata, and the expected structured output. Object metadata is represented as Full Natural Language Descriptions.

```
Your task is to identify object IDs referenced in the last user utterance.

### Instructions:
- You will be provided with a list of objects currently in the scene. Metadata for each object
    will also be provided.
- After that, you will receive the dialogue history.
- Your goal is to determine which objects, if any, are referred to in the last user utterance.
- Extract only the **object ID(s)** from the provided object list.
- **By default, assume no object is referenced** unless you are **highly confident**.
- Output object ID(s) **only when absolutely certain** that the last user utterance refers to
   them.
- If any objects are referenced in the last user utterance, output their ID(s) **between `<SOM
   >` and `<EOM>`**.
- **Use only the provided object IDs and metadata.** If an object is not listed, ignore it.
- Whenever an object ID appears inside `<SOM> ... <EOM>` in the dialogue history, it **
   explicitly marks the objects mentioned in that part of the conversation.**
- **Do NOT generate explanations, apologies, or extra text.**
- **Do NOT answer any questions in the conversation.**
- Format: `<SOM> ID1, ID2 <EOM>` (Example: `<SOM> 55, 85 <EOM>`).
- **If unsure or if the last user utterance is ambiguous, do NOT output any object IDs.**

### Current Objects in Scene:
ID: 0 : {"description": "grey Sofa from the brand Modern Arts. Made from leather. It has a
   customer rating of 3.1 out of 5. Priced at $399.00.", "location": "left"}
ID: 1 : {"description": "wooden EndTable from the brand Modern Arts. Made from wood. It has a
   customer rating of 3.3 out of 5. Priced at $399.00.", "location": "left"}
ID: 2 : {"description": "green Chair from the brand Home Store. Made from natural fibers. It
   has a customer rating of 2.9 out of 5. Priced at $399.00.", "location": "center"}
ID: 3 : {"description": "white Lamp from the brand Global Voyager. Made from metal. It has a
   customer rating of 3.1 out of 5. Priced at $349.00.", "location": "left"}
ID: 4 : {"description": "black EndTable from the brand North Lodge. Made from wood. It has a
   customer rating of 3.7 out of 5. Priced at $549.00.", "location": "left"}
ID: 5 : {"description": "grey AreaRug from the brand North Lodge. Made from wool. It has a
   customer rating of 3.8 out of 5. Priced at $499.00.", "location": "center"}
ID: 6 : {"description": "red Sofa from the brand River Chateau. Made from leather. It has a
   customer rating of 4.9 out of 5. Priced at $599.00.", "location": "left"}
ID: 7 : {"description": "white CouchChair from the brand Downtown Stylists. Made from leather.
    It has a customer rating of 4.4 out of 5. Priced at $499.00.", "location": "center"}
ID: 8 : {"description": "white AreaRug from the brand Home Store. Made from natural fibers. It
    has a customer rating of 3.6 out of 5. Priced at $449.00.", "location": "top-left"}

### Dialogue History:
User : Can you suggest a rug for me? System : You may like the geometric print one up front or
    the grey one in back by the partition. <SOM> <5>, <8> <EOM> User : How muh is the rug?
   System : Sorry, for which one? <SOM>  <EOM> User : Sorry I wasn't clear. I'd like to know
   about both rugs.

### Output:
Generated ID: <SOM>
```

Figure 7: Example prompt containing the full input configuration, including all object metadata and object references as provided to the model in the complete setting.

```
Your task is to identify object IDs referenced in the last user utterance.

### Instructions:
- You will be provided with a list of objects currently in the scene.
- After that, you will receive the dialogue history.
- Your goal is to determine which objects, if any, are referred to in the last user utterance.
- Extract only the **object ID(s)** from the provided object list.
- **By default, assume no object is referenced** unless you are **highly confident**.
- Output object ID(s) **only when absolutely certain** that the last user utterance refers to
    them.
- If any objects are referenced in the last user utterance, output their ID(s) **between `<SOM
    >` and `<EOM>`**.
- Whenever an object ID appears inside `<SOM> ... <EOM>` in the dialogue history, it **
    explicitly marks the objects mentioned in that part of the conversation.**
- **Do NOT generate explanations, apologies, or extra text.**
- **Do NOT answer any questions in the conversation.**
- Format: `<SOM> ID1, ID2 <EOM>` (Example: `<SOM> 55, 85 <EOM>`).
- **If unsure or if the last user utterance is ambiguous, do NOT output any object IDs.**

### Current Objects in Scene:
ID: 0, ID: 1, ID: 2, ID: 3, ID: 4, ID: 5, ID: 6, ID: 7, ID: 8

### Dialogue History:
User : Can you suggest a rug for me? System : You may like the geometric print one up front or
     the grey one in back by the partition. <SOM> <5>, <8> <EOM> User : How muh is the rug?
    System : Sorry, for which one? <SOM>  <EOM> User : Sorry I wasn't clear. I'd like to know
    about both rugs.

### Output:
Generated ID: <SOM>
```

Figure 8: Example prompt with object metadata omitted, retaining only the user utterance and object references to assess the model's reliance on attribute information.

```
Your task is to identify object IDs referenced in the last user utterance.

### Instructions:
- You will be provided with a list of objects currently in the scene. Metadata for each object
    will also be provided.
- After that, you will receive the dialogue history.
- Your goal is to determine which objects, if any, are referred to in the last user utterance.
- Extract only the **object ID(s)** from the provided object list.
- **By default, assume no object is referenced** unless you are **highly confident**.
- Output object ID(s) **only when absolutely certain** that the last user utterance refers to
   them.
- If any objects are referenced in the last user utterance, output their ID(s) **between `<SOM
   >` and `<EOM>`**.
- **Use only the provided object IDs and metadata.** If an object is not listed, ignore it.
- **Do NOT generate explanations, apologies, or extra text.**
- **Do NOT answer any questions in the conversation.**
- Format: `<SOM> ID1, ID2 <EOM>` (Example: `<SOM> 55, 85 <EOM>`).
- **If unsure or if the last user utterance is ambiguous, do NOT output any object IDs.**

### Current Objects in Scene:
ID: 0 : {"description": "grey Sofa from the brand Modern Arts. Made from leather. It has a
   customer rating of 3.1 out of 5. Priced at $399.00.", "location": "left"}
ID: 1 : {"description": "wooden EndTable from the brand Modern Arts. Made from wood. It has a
   customer rating of 3.3 out of 5. Priced at $399.00.", "location": "left"}
ID: 2 : {"description": "green Chair from the brand Home Store. Made from natural fibers. It
   has a customer rating of 2.9 out of 5. Priced at $399.00.", "location": "center"}
ID: 3 : {"description": "white Lamp from the brand Global Voyager. Made from metal. It has a
   customer rating of 3.1 out of 5. Priced at $349.00.", "location": "left"}
ID: 4 : {"description": "black EndTable from the brand North Lodge. Made from wood. It has a
   customer rating of 3.7 out of 5. Priced at $549.00.", "location": "left"}
ID: 5 : {"description": "grey AreaRug from the brand North Lodge. Made from wool. It has a
   customer rating of 3.8 out of 5. Priced at $499.00.", "location": "center"}
ID: 6 : {"description": "red Sofa from the brand River Chateau. Made from leather. It has a
   customer rating of 4.9 out of 5. Priced at $599.00.", "location": "left"}
ID: 7 : {"description": "white CouchChair from the brand Downtown Stylists. Made from leather.
    It has a customer rating of 4.4 out of 5. Priced at $499.00.", "location": "center"}
ID: 8 : {"description": "white AreaRug from the brand Home Store. Made from natural fibers. It
    has a customer rating of 3.6 out of 5. Priced at $449.00.", "location": "top-left"}

### Dialogue History:
User : Can you suggest a rug for me? System : You may like the geometric print one up front or
    the grey one in back by the partition. User : How muh is the rug? System : Sorry for
   which one?    User : Sorry I wasn't clear. I'd like to know about both rugs.

### Output:
Generated ID: <SOM>
```

Figure 9: Example prompt with object references omitted, retaining only the user utterance and metadata to assess the model's reliance on explicit object grounding cues.

```
ID: 7 : {
  "brand": "River Chateau",
  "color": "blue",
  "customerRating": 3.4,
  "materials": "wool",
  "price": "$249",
  "type": "AreaRug",
  "coordinates": {
    "x1_normalized": -0.0922,
    "x2_normalized": 0.2848,
    "y1_normalized": 0.4986,
    "y2_normalized": 0.4633,
    "area_normalized": 0.1054,
    "z_normalized": 0.606
  }
}
```

Figure 10: Structured representation of an object using raw normalized coordinates for spatial information.

```
ID: 7 : {
  "brand": "River Chateau",
  "color": "blue",
  "customerRating": 3.4,
  "materials": "wool",
  "price": "$249",
  "type": "AreaRug",
  "location": "top-right"
}
```

Figure 11: Structured representation of an object where spatial coordinates are replaced with natural language location descriptors (e.g., “top-right”).

```
ID: 7 : {
  "description": "blue AreaRug from the brand
       River Chateau. Made from wool. It has
      a customer rating of 3.4 out of 5.
      Priced at $249.00.",
  "location": "top-right"
}
```

Figure 12: Fully naturalized representation of an object, where all attributes including location are expressed as a fluent natural language description.