

INSIGHT: Structured Multimodal Knowledge Modeling for Vision-Guided Task-Oriented Dialogue

Xuelian Dong¹, Jiale Chen¹, Tianyong Hao^{1*}

¹School of Computer Science, South China Normal University, China

{xldong, jlchen, haoty}@m.scnu.edu.cn

Abstract

In real-world task-oriented dialogue (TOD), users often express their intents through both text and images, with images providing essential cues for fine-grained requirements. Addressing such scenarios demands effective cross-modal semantic modeling and precise knowledge retrieval. However, existing TOD methods primarily focus on text and assume one-to-one attribute mappings, failing to reflect the multimodal nature of real-world information. To overcome these limitations, this paper proposes an end-to-end vision-guided retrieval-generation framework named INSIGHT. It introduces a progressive dynamic fusion mechanism that performs layer-wise cross-modal interaction to achieve robust multimodal dialogue understanding. In addition, a graph-structured entity modeling strategy explicitly models one-to-many image attribute mappings through distributed latent semantic learning, thereby preserving visual diversity. Finally, knowledge retrieval and response generation are jointly optimized in an end-to-end manner. To evaluate real-world scenarios, we construct VISTA, a new multimodal TOD dataset that features image-driven multi-turn dialogues and supports one-to-many entity attribute mappings. Experimental results on two multimodal datasets demonstrate that INSIGHT consistently outperforms existing methods across various backbone models¹.

1 Introduction

Task-Oriented Dialogue (TOD) systems aim to generate informative and natural responses by understanding dialogue context and retrieving external Knowledge Bases (KBs) (Rao et al., 2026; Gao et al., 2024). In recent years, such systems have been widely applied in various scenarios, includ-

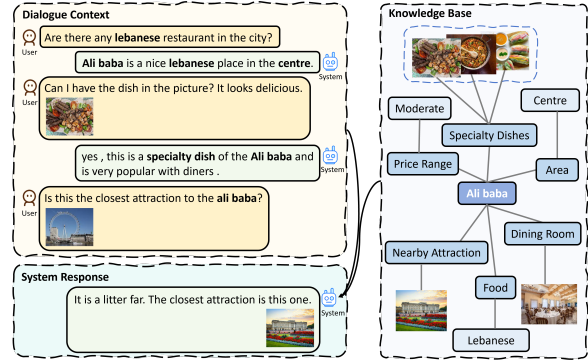


Figure 1: A restaurant reservation example in our multimodal task-oriented dialogue system.

ing restaurant booking, weather inquiry, and travel planning (Shi et al., 2023; Ding et al., 2024).

However, existing studies mainly focus on a text-only paradigm, which fails to capture the multimodal characteristics of real-world interactions (Saab et al., 2026; Yang et al., 2021; Kottur et al., 2021). In scenarios with more precise requirements, users often express their needs through a combination of text and images (Feng et al., 2023; Chen et al., 2023). In particular, images play an indispensable role in capturing visual nuances and contextual cues that are difficult to express through text alone. Moreover, responding to user-specific image-text inquiries often necessitates retrieving the most relevant knowledge from real-world data that involves one-to-one or one-to-many mappings. As illustrated in Figure 1, a user sends an image of a dish and inquires whether the dish is available at a given restaurant. Answering such a user question requires not only accurate understanding of the dish image but also deep semantic retrieval among multiple similar dish images under the “*specialty dishes*” attribute to generate a correct response.

Despite the growing demand for multimodal dialogue, the incorporation of visual information into TODs has received limited attention, with only a handful of studies addressing this issue (Meng

*Corresponding author

¹Our dataset and code are publicly available at: <https://github.com/Libv-Team/INSIGHT>.

et al., 2026; Kong et al., 2024). However, these few studies face two main limitations. First, task modeling driven by multimodal questions is generally lacking. Most related studies focus on textual interactions, merely introducing images as auxiliary attributes in the KB (Ma et al., 2022; Liu et al., 2024). Second, current TOD systems lack mechanisms to model one-to-many mappings within entity attributes (Huang et al., 2024). This limitation primarily arises from the design of existing multimodal KBs, which support only one-to-one mappings, where each attribute corresponds to a single text value or a single image (Dong et al., 2025).

Therefore, advancing TOD systems toward realistic multimodal scenarios remains highly challenging. First, user intents are often conveyed through images rather than explicit textual slots, making intent understanding and task grounding more ambiguous than in text-only settings. Second, one-to-many multimodal attribute mappings complicate knowledge retrieval. Multiple images under the same attribute are often visually similar but semantically distinct, making it difficult to identify the exact evidence required by the user. Flattening them into one-to-one pairs removes the shared attribute constraint and further increases retrieval ambiguity. Finally, most existing TOD systems decouple knowledge retrieval from response generation, such independent optimization amplifies error propagation and imposes stringent requirements on retrieval accuracy (Chen et al., 2025c; Dong and Chen, 2023; Chen et al., 2024).

To address these challenges, an end-to-end vision-guided retrieval-generation framework named INSIGHT is proposed. It explicitly organizes the multimodal KB as graph-structured entities, in which attribute values are hierarchically aggregated into entity representations. Moreover, we introduce a progressive dynamic fusion mechanism that aligns textual and visual semantics for robust dialogue understanding. To enable fine-grained multimodal knowledge grounding, we introduce a hierarchical semantic-perceptual retrieval strategy that combines entity-level semantic localization with image-level perceptual discrimination, and further employs differentiable soft retrieval to jointly optimize retrieval and generation end-to-end. To better reflect realistic multimodal dialogue scenarios, we construct VISTA, a new benchmark featuring image-driven user questions and one-to-many multimodal attribute mappings. Experimental results on VISTA and another multi-

modal TOD benchmark demonstrate that INSIGHT consistently achieves superior performance across different backbone models. In summary, our main contributions are as follows:

- An end-to-end vision-guided retrieval-generation framework INSIGHT is proposed, which unifies cross-modal alignment, multimodal retrieval, and generation in an end-to-end manner, thereby improving response quality.
- A graph-structured entity modeling strategy is proposed for multimodal KBs with one-to-many attribute mappings, where visual diversity and structural uncertainty are preserved through distributed latent semantic modeling.
- A novel multimodal dataset named VISTA is constructed, which is the first to feature image-driven user questions and one-to-many attribute-value mappings, providing a more challenging and realistic benchmark.

2 The Method

2.1 Problem Definition

In the multimodal TOD setting, the dialogue context is denoted as $C = [H, Q]$, where H is the dialogue history and $Q = (U_n^{text}, U_n^{image})$ is the current user question. The system response is denoted as $R = (R_n^{text}, R_n^{image})$, where the visual modality is optional and may be $\emptyset$. To explicitly model the structured organization of multimodal knowledge, the KB is represented as a collection of entity-centered knowledge graphs $K = \{G_1, G_2, \dots, G_T\}$. Each graph G_i is centered on an entity node e_i , where each attribute $a_{i,m}$ is modeled as a child node and each attribute value is represented as a leaf node connected to its corresponding attribute node. For text-valued attributes, the mapping is one-to-one, i.e., each attribute node is associated with a single textual leaf node $v_{i,m}^{text}$. For image-valued attributes, the mapping is one-to-many, i.e., each attribute node may correspond to a set of visually similar image leaf nodes $v_{i,m}^{image} = \{I_{i,m}^1, I_{i,m}^2, \dots, I_{i,m}^J\}$.

2.2 Progressive Dynamic Fusion

To better model multimodal dialogue context, we propose a progressive dynamic fusion mechanism. Unlike late-fusion methods that combine multimodal representations only at the final stage, it

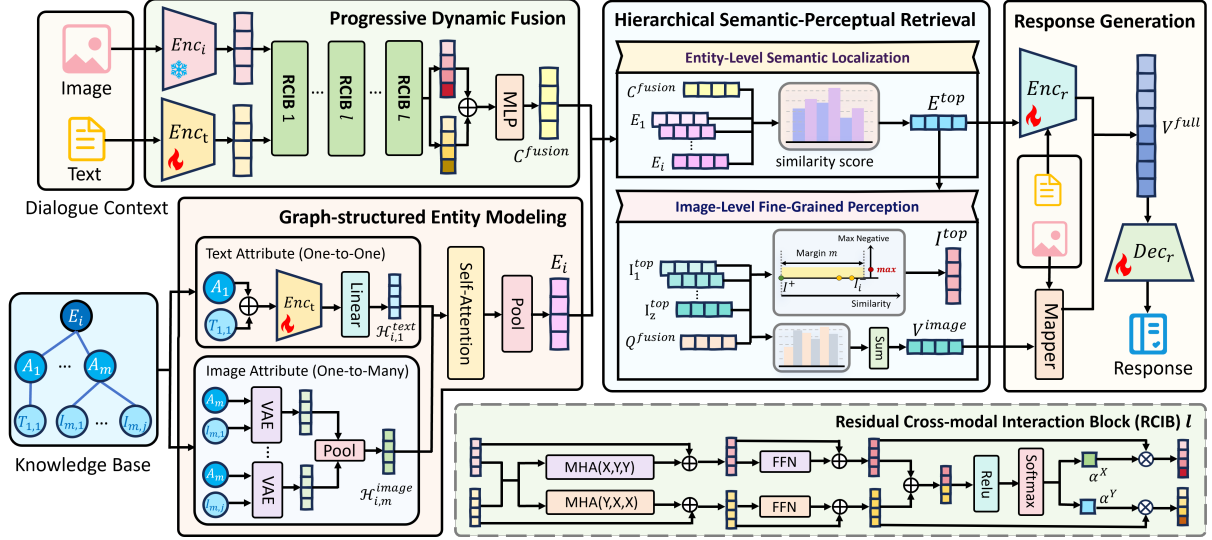


Figure 2: The overall architecture of INSIGHT, which consists of four components: progressive dynamic fusion, graph-structured entity modeling, hierarchical semantic-perceptual retrieval, and response generation.

performs cross-modal refinement at every fusion layer and dynamically adjusts modality contributions according to the semantic context at that layer. This design enables the two modalities to interact progressively across layers while mitigating the information compression caused by one-shot fusion.

Given a textual input T and a visual input I , we obtain their modality-specific representations $h_0^{text} = Enc_t(T)$ and $h_0^{image} = Enc_i(I)$ through a text encoder and an image encoder, respectively. Building upon these initial representations, we apply the progressive dynamic fusion in a layer-wise manner. At the l -th fusion layer, the textual and visual states are reciprocally refined as

$$\tilde{h}_l^{text} = \mathcal{R}_l^t(h_{l-1}^{text}, h_{l-1}^{image}), \quad (1)$$

$$\tilde{h}_l^{image} = \mathcal{R}_l^i(h_{l-1}^{image}, h_{l-1}^{text}), \quad (2)$$

where $\mathcal{R}_l(\cdot)$ is a standard residual cross-modal interaction block. Each block is defined as

$$S_l = \text{LN}(X + \text{MHA}_l(X, Y, Y)), \quad (3)$$

$$\mathcal{R}_l(X, Y) = \text{LN}(S_l + \text{FFN}_l(S_l)), \quad (4)$$

where $\text{MHA}_l(\cdot)$ denotes multi-head cross-attention, $\text{FFN}_l(\cdot)$ is a position-wise feed-forward network, and $\text{LN}(\cdot)$ denotes LayerNorm.

To adaptively calibrate modality contributions at each layer, we compute layer-wise gating weights from the refined states:

$$[\alpha_l^{text}, \alpha_l^{image}] = \text{Softmax}\left(\mathcal{G}_l([\tilde{h}_l^{text}; \tilde{h}_l^{image}])\right), \quad (5)$$

$$\mathcal{G}_l(x) = W_{g,2} \text{ReLU}(W_{g,1}x). \quad (6)$$

The modality-specific states are then updated as:

$$h_l^{text} = \eta \alpha_l^{text} \odot \tilde{h}_l^{text}, \quad (7)$$

$$h_l^{image} = \eta \alpha_l^{image} \odot \tilde{h}_l^{image}, \quad (8)$$

where $\odot$ denotes element-wise multiplication. The scaling factor $\eta = 2$ compensates for the average attenuation introduced by softmax normalization over the two modalities.

After $L = 3$ layers of progressive interaction, we obtain the final modality representations h_L^{text} and h_L^{image} . The final multimodal representation is then formed as

$$h^{fusion} = \phi([h_L^{text}; h_L^{image}]), \quad (9)$$

where $\phi(\cdot)$ is a two-layer MLP with LayerNorm and GELU that projects the concatenated multimodal features into a unified semantic space.

Application. For the multimodal dialogue context C , the progressive dynamic fusion mechanism $\mathcal{F}_{PDF}(\cdot)$ aligns textual dialogue context C^{text} with the image utterance U_n^{image} in the user question Q , generating the fused dialogue context C^{fusion} .

$$C^{fusion} = \mathcal{F}_{PDF}(C^{text}, U_n^{image}) \quad (10)$$

In addition, the fused user question is represented as Q^{fusion} :

$$Q^{fusion} = \mathcal{F}_{PDF}(U_n^{text}, U_n^{image}). \quad (11)$$

2.3 Graph-structured Entity Modeling

In multimodal TOD systems, integrating external knowledge is challenging due to structural complexity and one-to-many modality mappings. To explicitly model this feature, we propose a graph-structured entity modeling strategy. For a given entity graph G_i , our strategy operates in a bottom-up manner, aggregating semantic information from heterogeneous leaf nodes (attribute values) to the central root node (entity), thereby yielding a comprehensive entity representation.

For standard text attributes with a one-to-one mapping, we simply extract the joint features of the concatenated attribute name and value using a text encoder Enc_t , and project them into a shared latent space via a linear layer to obtain the node feature. Formally,

$$\mathcal{H}_{i,m}^{text} = W_t Enc_t([a_{i,m}; v_{i,m}^{text}]) + b_t. \quad (12)$$

However, for visual attributes exhibiting a one-to-many relationship, naive average pooling severely dilutes the visual variance and structural uncertainty introduced by diverse images. To effectively encapsulate this variance, we design an attribute Variational AutoEncoder (VAE). We first extract the textual feature of the attribute name and the visual feature of the j -th image. The concatenated representation is formed as a joint hidden state $h_{i,m}^j = [Enc_t(a_{i,m}); Enc_i(I_{i,m}^j)]$. Instead of a deterministic encoding, we map each image to a multi-dimensional Gaussian distribution by predicting its mean μ and log-variance $\log \sigma^2$:

$$\mu_{i,m}^j = W_\mu h_{i,m}^j, \quad \log(\sigma_{i,m}^j)^2 = W_\sigma h_{i,m}^j. \quad (13)$$

Using the reparameterization trick, we sample robust, noise-resilient latent variables by injecting a standard Gaussian noise vector ϵ , where $\epsilon \sim \mathcal{N}(0, I)$ indicates that each dimension is independently sampled from a normal distribution with zero mean and unit variance:

$$\tilde{\mathcal{H}}_{i,m}^{j,image} = \mu_{i,m}^j + \epsilon \odot \sigma_{i,m}^j, \quad \epsilon \sim \mathcal{N}(0, I). \quad (14)$$

The node representation for the multi-image attribute is then computed by mean-pooling the sampled states over all J images:

$$\mathcal{H}_{i,m}^{image} = \frac{1}{J} \sum_{j=1}^J \tilde{\mathcal{H}}_{i,m}^{j,image}. \quad (15)$$

To regularize the latent space during training, we minimize the Kullback-Leibler (KL) divergence

between the learned distributions and the standard normal prior $\mathcal{N}(0, I)$:

$$\mathcal{L}_{KL} = -\frac{1}{2} \sum_{j=1}^J \sum_{d=1}^D \left(1 + \log(\sigma_d^j)^2 - (\mu_d^j)^2 - (\sigma_d^j)^2 \right), \quad (16)$$

where D is the hidden dimensionality, and μ_d^j and σ_d^j denote the d -th dimension of the predicted mean and standard deviation for the j -th image.

Finally, to capture semantic dependencies among the M attributes, we stack their representations into an attribute-node sequence $\mathcal{H}_i^A = [\mathcal{H}_{i,1}; \dots; \mathcal{H}_{i,M}]$ and process it with a multi-head self-attention module. This allows each attribute node to attend to all others and encode inter-attribute dependencies within the entity. These contextually enriched attribute nodes $\hat{\mathcal{H}}_i^A$ are then aggregated back to the central entity node to yield the final dense representation:

$$E_i = \text{LN} \left(\frac{1}{M} \sum_{m=1}^M \hat{\mathcal{H}}_i^A \right). \quad (17)$$

Overall, our graph-structured entity modeling strategy brings two main advantages. It preserves the diversity of one-to-many image mappings through distributed variance modeling, thereby improving robustness to noisy or conflicting visual signals. Meanwhile, by projecting textual and visual branches into a shared semantic space, it enables self-attention to capture richer cross-modal dependencies than flat concatenation.

2.4 Hierarchical Semantic-Perceptual Retrieval

To achieve reliable localization of multimodal knowledge, we propose a hierarchical semantic-perceptual retrieval mechanism that realizes robust retrieval through a coarse-to-fine hierarchical constraint. Specifically, the mechanism first performs dialogue-driven entity-level semantic localization, and then conducts image-level perceptual retrieval under entity constraints, thereby effectively bridging abstract semantic understanding with fine-grained perceptual discrimination.

Entity-Level Semantic Localization. In the first stage of retrieval, we use the fused dialogue context C^{fusion} as a semantic query to perform holistic semantic localization over candidate entities in the knowledge base K . The objective of this stage is to accurately identify the top-1 entity whose semantics are most consistent with the current dialogue context from the global entity set

$\mathcal{E} = \{E_1, \dots, E_T\}$. For each dialogue turn, we apply a standard InfoNCE contrastive loss to pull the C^{fusion} towards the ground-truth positive entity E^+ while pushing apart negative entities:

$$\mathcal{L}_{ent} = -\log \frac{\exp(S(C^{fusion}, E^+))}{\sum_{t=1}^T \exp(S(C^{fusion}, E_t))}, \quad (18)$$

where $S(C, E) = \cos(C, E)/\tau$, and τ is a temperature hyperparameter.

After optimization, the entity achieving the highest similarity score with C^{fusion} is selected as the top-1 most relevant entity E^{top} :

$$E^{top} = \arg \max_{E_t \in \mathcal{E}} \cos(C^{fusion}, E_t). \quad (19)$$

Image-Level Fine-Grained Perception. At this stage, all candidate images are drawn from E^{top} . While their overall semantics are similar, they often exhibit critical differences in attribute details or local visual features. Similarly, the fused user question Q^{fusion} is employed to perform fine-grained retrieval over the entity-associated image set $\mathcal{I} = \{I_1, \dots, I_Z\}$. Beyond the standard InfoNCE objective $\mathcal{L}_{info}$, we further introduce a maximum-negative-based margin constraint a , which explicitly enforces a minimum discriminative distance between positive and negative samples. This design alleviates the hard-negative interference caused by the high visual similarity among images associated with the same entity. Formally,

$$\mathcal{G} = a + \max_i S(Q^{fusion}, I_i) - S(Q^{fusion}, I^+), \quad (20)$$

$$\mathcal{L}_{img} = \mathcal{L}_{info} + \max(0, \mathcal{G}). \quad (21)$$

During inference, the image ranked as top-1 based on its similarity score with Q^{fusion} is identified as the most relevant image I^{top} :

$$I^{top} = \arg \max_{I_z \in \mathcal{I}} \cos(Q^{fusion}, I_z). \quad (22)$$

The most relevant image retrieved in this way is directly used as the final visual response R_n^{image} .

Differentiable Soft Retrieval for Generation. Conventional retrieval simply adopts the top-1 discrete image as the output, which breaks the gradient flow for downstream generation. To address this, we compute a normalized probability distribution over all candidate images, and perform a weighted aggregation to obtain a soft representation V^{img} :

$$V^{img} = \sum_{z=1}^Z \frac{\exp(S(Q^{fusion}, I_z))}{\sum_{j=1}^Z \exp(S(Q^{fusion}, I_j))} I_z. \quad (23)$$

This soft visual representation V^{img} is subsequently mapped and injected into the generative model, bridging the retrieval and generation phases in an end-to-end differentiable manner.

2.5 Response Generation

After retrieving the structured multimodal knowledge, the final step is to generate an accurate and context-aware system response. To leverage the reasoning capabilities of the generation model while seamlessly accommodating visual information, we project all selected heterogeneous features into a unified continuous embedding sequence. Specifically, we utilize a trainable modality continuous mapper $\mathcal{M}$ to project visual features into the same embedding space as the generative model’s embeddings. The comprehensive multimodal input sequence V_{full} is constructed by directly concatenating the dialogue context representations $V_{ctx} = [Enc_r(C^{text}); \mathcal{M}(U_n^{image})]$ and the retrieved knowledge representations $V_{ent} = [Enc_r(E^{top}); \mathcal{M}(V^{img})]$, where Enc_r is the response encoder. Formally,

$$V_{full} = [V_{ctx}; V_{ent}]. \quad (24)$$

Conditioned on V_{full} , the generative model autoregressively predicts the target textual response $R_n^{text} = \{r_1, \dots, r_Y\}$. We apply the standard negative log-likelihood loss for generation:

$$\mathcal{L}_{gen} = -\sum_{y=1}^Y \log P(r_y | r_{<y}, V_{full}). \quad (25)$$

Joint Training Objective. A paramount advantage of our differentiable architecture is that it permits end-to-end joint training. Rather than optimizing retrieval and generation in isolated phases, our framework propagates gradients from the generation loss back to the multimodal encoders. The overall objective function is a weighted sum of four corresponding task losses:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{gen} + \lambda_2 \mathcal{L}_{ent} + \lambda_3 \mathcal{L}_{img} + \lambda_4 \mathcal{L}_{KL}, \quad (26)$$

where $\lambda_1 = 0.4$, $\lambda_2 = 0.3$, $\lambda_3 = 0.3$, and $\lambda_4 = 0.01$ are scaling hyperparameters.

3 The VISTA Dataset

To advance research on multimodal task-oriented dialogue, we construct VISTA, a novel multi-domain image-guided dataset covering restaurant reservation, hotel booking, and travel assistance. Unlike prior multimodal TOD datasets, VISTA

Dataset	Dialogue Context	KB	Response	Image-Driven User Questions	Multimodal One-to-Many Attribute Mapping
Camrest (Wen et al., 2017)	Text	Text	Text	●	●
SMD (Eric et al., 2017)	Text	Text	Text	●	●
MWOZ (Budzianowski et al., 2018)	Text	Text	Text	●	●
MMDialKB (Yang et al., 2021)	Text	Text + Image	Text	●	●
MuCoK (Dong et al., 2025)	Text + Image	Text + Image	Text + Image	●	●
VISTA (Ours)	Text + Image	Text + Image	Text + Image	●	●

Table 1: Comparison between our VISTA dataset and five commonly used datasets in TOD systems.

Statistic	Train	Valid	Test
Dialogues	642	275	276
Utterances	8159	3419	3406
Avg length of utterances	13.57	14.69	14.61
Utterances with entities	4012	1135	1124

Table 2: Detailed Statistics of the VISTA Dataset.

adopts an image-driven interaction setting in which user intents are primarily conveyed through images, with text providing complementary information. In addition, the knowledge base contains multiple visually similar images grouped under a single image attribute to form a one-to-many mapping relationship. This design highlights two key challenges: TOD systems must correctly interpret the semantics of user-provided images to satisfy task requirements, and further retrieve the most relevant image from sets of visually similar candidates based on joint text-image semantics. By explicitly modeling this scenario, VISTA provides the first benchmark that integrates image-centered user intent with fine-grained image retrieval for generating image-text responses, thereby offering a valuable resource in multimodal TOD research.

3.1 Data Collection Process

Inspired by the pure-text Camrest dataset (Wen et al., 2017) and the multimodal MMDialKB dataset (Yang et al., 2021), we construct VISTA by extending conventional TOD data organization to support image-text interactions. Each dialogue turn consists of an image-text user question and an image-text system response grounded in a multimodal knowledge base containing both textual and visual information. To construct the visual component of the knowledge base, candidate images were manually collected through web search engines. Images with watermarks or low resolution were removed, and multiple visually similar images were grouped under the same image-related attribute to establish the one-to-many mapping. Subsequently, multimodal dialogues were manually collected from audios and reviewed to minimize issues such as ambiguous context, entity mismatches, and spelling errors. The user ques-

tions are mostly image-driven, with appropriate textual supplements. Importantly, the image-text mixed system responses include detailed textual descriptions of the image content, which increases the complexity of understanding the image semantics.

To better simulate real-world multimodal dialogue scenarios, the images annotated in user question are redrawn using the stable-diffusion-xl-refiner-1.0 (Rombach et al., 2022) model based on the original images from the knowledge base. This ensures that the user question images retained the main features of the original images while presenting different perspectives such as top view, frontal view, or rear view. The redrawn images are manually checked and corrected, with non-compliant abstract images removed and regenerated. Finally, we invited three experts in the field of dialogue generation to conduct an additional round of review and refinement on the annotated data. This process helped eliminate issues such as grammatical errors and unclear semantic expressions, ensuring the high quality and usability of the VISTA dataset. Examples of the VISTA dataset are provided in Appendix G.

3.2 Dataset Statistics

To highlight the unique properties of VISTA, we compare it with five commonly used task-oriented dialogue datasets, as shown in Table 1. It can be observed that Camrest (Wen et al., 2017), SMD (Eric et al., 2017), and MultiWOZ (Budzianowski et al., 2018) are all single-modal datasets composed entirely of textual content. Although MMDialKB (Yang et al., 2021) introduces a multimodal knowledge base, both its dialogue context and responses remain in pure textual form. The MuCoK dataset (Dong et al., 2025) includes text-image dialogue contexts, knowledge base, and responses; however, it lacks image-driven user questions and does not support multiple images as a single attribute value of an entity. In contrast, our VISTA jointly supports all these properties. VISTA contains 14,984 utterances and 5,307 images. Detailed dataset statistics are shown in Table 2.

Dataset	Type	Backbone	BLEU	Entity Precision	Entity Recall	Entity F1	Entity Accuracy	Image Accuracy
MuCoK	RAG + MLLM	Qwen2.5-VL-7B [†]	0.52	33.47	50.43	40.24	25.42*	64.77*
		Qwen2.5-VL-7B [‡]	0.49	51.57	34.60	41.41	25.42*	64.77*
		Gemma3-4B [†]	0.38	42.62	51.29	46.56	25.42*	64.77*
		Gemma3-4B [‡]	2.52	43.07	50.70	46.58	25.42*	64.77*
		Qwen3-VL-8B [†]	0.37	32.75	51.17	39.94	25.42*	64.77*
		Qwen3-VL-8B [‡]	0.64	32.37	41.70	36.45	25.42*	64.77*
	RAG + LM/LLM	FlanT5-Base	6.21	55.87	30.80	39.71	25.42*	64.77*
		FlanT5-Large	10.31	60.63	47.54	53.29	25.42*	64.77*
		Qwen2.5-7B	8.70	65.83	35.09	45.78	25.42*	64.77*
		LLaMA3-8B	10.29	71.05	35.72	47.54	25.42*	64.77*
		Qwen3-8B	8.69	69.71	36.57	47.97	25.42*	64.77*
	INSIGHT (Ours)	FlanT5-Base	<u>8.89</u>	66.86	67.96	67.40	70.84	86.32
		FlanT5-Large	15.88	84.97	78.47	81.59	73.29	83.23
		Qwen2.5-7B	5.81	<u>78.54</u>	78.29	78.41	63.23	<u>87.10</u>
		LLaMA3-8B	6.22	78.12	<u>82.39</u>	<u>80.19</u>	72.90	86.84
		Qwen3-8B	6.44	74.79	83.16	78.75	73.29	88.52
VISTA	RAG + MLLM	Qwen2.5-VL-7B [†]	0.36	31.21	46.67	37.41	18.46*	50.98*
		Qwen2.5-VL-7B [‡]	0.50	35.53	30.45	32.80	18.46*	50.98*
		Gemma3-4B [†]	0.28	39.43	45.76	42.36	18.46*	50.98*
		Gemma3-4B [‡]	2.88	38.26	51.65	43.96	18.46*	50.98*
		Qwen3-VL-8B [†]	0.30	32.64	50.46	39.64	18.46*	50.98*
		Qwen3-VL-8B [‡]	1.17	49.28	50.84	50.05	18.46*	50.98*
	RAG + LM/LLM	FlanT5-Base	6.28	59.70	34.80	43.97	18.46*	50.98*
		FlanT5-Large	9.76	66.85	44.25	53.25	18.46*	50.98*
		Qwen2.5-7B	6.77	42.51	51.15	46.43	18.46*	50.98*
		LLaMA3-8B	5.41	46.58	53.39	49.75	18.46*	50.98*
		Qwen3-8B	4.98	47.33	56.25	51.41	18.46*	50.98*
	INSIGHT (Ours)	FlanT5-Base	<u>9.77</u>	62.91	76.60	69.09	74.90	<u>86.03</u>
		FlanT5-Large	16.40	83.46	81.81	82.62	<u>75.39</u>	85.06
		Qwen2.5-7B	5.64	60.65	69.36	64.71	76.66	86.82
		LLaMA3-8B	4.86	<u>77.16</u>	<u>80.44</u>	<u>78.77</u>	70.90	82.32
		Qwen3-8B	6.23	63.42	72.41	67.61	73.73	85.45

Table 3: Overall results on the MuCoK and our VISTA datasets. Statistically significant improvement over the best-performing baseline ($p < 0.05$). The best results are highlighted in bold, and the second-best results are underlined. [†]: zero-shot settings. [‡]: Settings for LoRA fine-tuning using cross-entropy loss. *: Standard contrastive learning is used to train the CLIP model for entity and image retrieval.

4 Experiments and Results

4.1 Settings

Evaluation Metrics. To comprehensively evaluate the performance of multimodal TOD, six metrics are adopted: BLEU (Papineni et al., 2002), Entity F1 (Eric et al., 2017), Entity Precision, Entity Recall, Entity Accuracy, and Image Accuracy. Details of these metrics are provided in Appendix H.

Implementation Details. INSIGHT is built upon FlanT5-Large and CLIP. CLIP is used as both the text and image encoder, while FlanT5-Large serves as the response encoder and decoder. INSIGHT is trained using the AdamW optimizer, with learning rates of 1×10^{-4} for the backbone models and 1×10^{-3} for the remaining trainable modules. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU with 24 GB memory.

4.2 The Overall Results

Table 3 summarizes the overall results on the MuCoK and VISTA datasets. The experimental results show that INSIGHT consistently achieves superior

performance across both datasets, with notable improvements in Entity F1, Entity Accuracy, and Image Accuracy, demonstrating its consistent effectiveness across datasets and backbone models. Specifically, on the MuCoK dataset, INSIGHT improves entity-related metrics across different backbone models, while Image Accuracy increases to 83%–88%, indicating its effectiveness in both entity recognition and cross-modal alignment. On the more challenging VISTA dataset, these advantages become even more pronounced. Baseline methods exhibit larger performance drops in Entity Accuracy and Image Accuracy, whereas INSIGHT maintains stable improvements, demonstrating its effectiveness in complex multimodal scenarios.

4.3 Impact of Visual Semantics on Retrieval

To evaluate the impact of correctly interpreting users’ visual semantics on knowledge retrieval performance, we visualize the confusion matrices. As shown in Figure 3 (a) and (c), after incorporating visual semantics, the confusion matrices exhibit

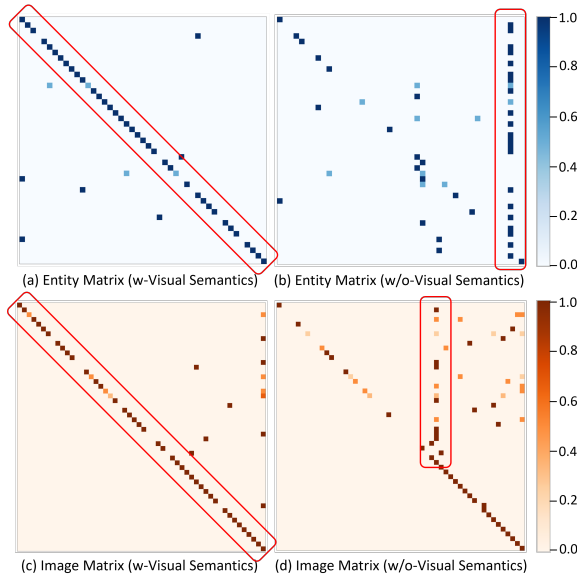


Figure 3: Impact of visual semantics on entity and image retrieval in the VISTA dataset.

strong concentration along the diagonal, indicating that the model can more accurately capture user intentions and retrieve the most relevant entities and images. In contrast, Figure 3 (b) and (d) show that removing visual input reduces the concentration of diagonal elements and increases retrieval errors. In particular, entities and images are frequently misclassified into the null category (highlighted by the red boxes). Moreover, without visual semantics, image retrieval still performs slightly better than entity retrieval. This phenomenon may be attributed to the differentiable soft retrieval, which enables unified optimization between retrieval and generation, allowing response quality to provide backward optimization signals for image retrieval. Overall, these results verify the necessity of correctly understanding visual semantics within dialogue contexts for multimodal TOD systems.

4.4 One-to-Many Modeling under Visual Diversity

To analyze the impact of one-to-many attribute mapping, we divide the samples into Easy, Medium, and Hard intervals based on the average image similarity among attributes within each entity, where higher similarity indicates greater visual confusion. As shown in Figure 4, directly modeling one-to-many mappings outperforms decomposing them into multiple one-to-one mappings across all metrics and similarity intervals. This indicates that preserving the one-to-many mapping structure helps the model capture relationships among visu-

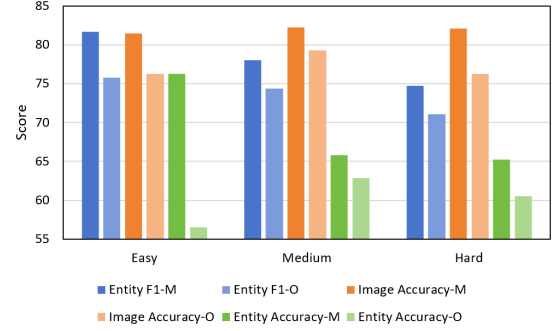


Figure 4: Comparison of one-to-many mapping under different difficulty intervals. -M denotes directly modeling one-to-many attribute mappings, while -O denotes decomposition into multiple one-to-one mappings.

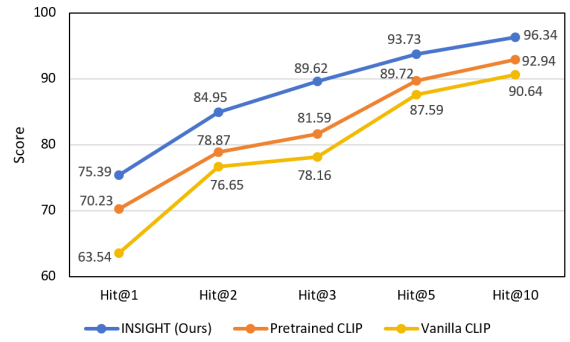


Figure 5: Entity hit rate comparison on the VISTA dataset.

ally similar attribute values, while decomposition weakens such structural connections and degrades knowledge retrieval. Overall, our graph-structured knowledge modeling better alleviates attribute confusion and maintains more stable performance in visually challenging scenarios.

4.5 Entity Hit Rate Analysis

Figure 5 presents the entity hit rate results under different Hit@ k settings. INSIGHT consistently achieves the best performance across all settings, especially under the strict Hit@1 metric, indicating its stronger ability to rank the correct entity at the top. As k increases, the hit rates of the methods using concatenated features from contrastive-loss-pretrained CLIP or vanilla CLIP also improve steadily, indicating that larger candidate sets better cover the correct entity. However, their weaker performance under low- k settings suggests limited ability to prioritize key entities. Overall, the advantage of INSIGHT is more pronounced under low- k settings, demonstrating its effectiveness in cross-modal alignment and fine-grained entity retrieval.

Method	BLEU	Entity F1	Image Accuracy
Step-by-Step Training Strategy	6.08	57.23	78.91
Retrieval + Generation Training Strategy	7.03	58.24	79.49
Unified Optimization Strategy (Ours)	16.40	82.62	85.06

Table 4: Performance comparison of different training strategies on the VISTA dataset.

Method	BLEU	Entity F1	Image Accuracy
INSIGHT (Ours)	16.40	82.62	85.06
w/o Progressive Dynamic Fusion	11.50 (4.90↓)	78.86 (3.76↓)	83.69 (1.37↓)
w/o Graph-structured Entity Modeling	10.34 (6.06↓)	75.62 (7.00↓)	76.07 (8.99↓)
w/o Differentiable Soft Retrieval	11.63 (4.77↓)	77.43 (5.19↓)	82.03 (3.03↓)
w/o Response Encoder	0.28 (16.12↓)	46.19 (36.43↓)	84.57 (0.49↓)

Table 5: Ablation study of four modules on the VISTA dataset where “w/o” means without.

4.6 Impact of Training Strategies

Table 4 compares the performance of different training strategies on the VISTA dataset. The unified optimization strategy achieves the best results across all metrics. Specifically, the step-by-step training strategy optimizes entity retrieval, image retrieval, and response generation separately, limiting the interaction among subtasks and resulting in inferior performance. The retrieval + generation training strategy only marginally improves performance on this basis. In contrast, the unified optimization strategy employs differentiable retrieval for end-to-end joint optimization, enabling different tasks to mutually reinforce each other.

4.7 Ablation Study

Table 5 presents the ablation results. Removing any component consistently degrades performance, demonstrating the effectiveness of each module in INSIGHT. In particular, removing the response encoder considerably degrades BLEU and Entity F1 while having only a minor effect on Image Accuracy, indicating that it primarily plays a critical role in response generation and entity-aware language modeling rather than image selection. Among the four core modules, graph-structured entity modeling exhibits the most pronounced and consistent impact across all evaluation metrics. Its removal notably weakens both entity prediction and image selection, highlighting the importance of explicitly modeling structural one-to-many relations among entities for effective cross-modal reasoning. Progressive dynamic fusion and differentiable soft retrieval also make clear contributions, with the former facilitating effective multimodal interaction and the latter supporting differentiable retrieval and end-to-end optimization.

5 Conclusion

This paper presents INSIGHT, an end-to-end vision-guided retrieval-generation framework. By introducing progressive dynamic fusion for cross-modal semantic interaction and graph-structured knowledge modeling for one-to-many multimodal attribute representation, INSIGHT improves multimodal dialogue understanding, knowledge grounding, and response generation. Moreover, we construct VISTA, a new benchmark that better reflects realistic vision-guided dialogue scenarios and fine-grained multimodal retrieval challenges. Experimental results on two datasets show that INSIGHT adapts well to complex multimodal scenarios across different backbones. Overall, INSIGHT and VISTA jointly advance TOD research toward more realistic image-text interaction settings.

Acknowledgements

The work is supported by grants from National Natural Science Foundation of China (No. 62372189).

Limitations

Although INSIGHT achieves promising results in image-guided multimodal TOD, it still has some limitations. First, this work mainly focuses on image-text interaction, while real-world dialogue scenarios may involve more diverse modalities, such as speech, video, and user behavior signals. These modalities often contain stronger temporal and dynamic information, which could be explored in future work to support more natural human-computer interaction. In addition, the current evaluation mainly relies on automatic metrics and limited human assessment, which may not fully capture user satisfaction, dialogue coherence, and long-term task completion. Future work could introduce interactive evaluation, real-user testing, and error recovery analysis to better assess the practical effectiveness of multimodal TOD systems.

References

- Hanbo Bi, Zhiqiang Yuan, Zexi Jia, Jiawei Zhang, Chongyang Li, Peixiang Luo, Ying Deng, Xiaoyue Duan, and Jinchao Zhang. 2026. F2rvlm: Boosting fine-grained fragment retrieval for multi-modal long-form dialogue with vision language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 14493–14501.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ra-

- madan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.
- Youngjae Chang and Youngjoong Ko. 2023. Two-step masked language model for domain-adapting multi-modal task-oriented dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2938–2943.
- Jiale Chen, Xuelian Dong, Shunhao Li, Heng Weng, and Tianyong Hao. 2025a. Multi-view knowledge retrieval with a new memory tracking mechanism for task-oriented dialogue. *IEEE Transactions on Consumer Electronics*.
- Jiale Chen, Xuelian Dong, Wenxiu Xie, Tao Gong, Fu Lee Wang, and Tianyong Hao. 2025b. Span attention for entity-consistent task-oriented dialogue response generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Jiale Chen, Xuelian Dong, Wenxiu Xie, Ru Peng, Kun Zeng, and Tianyong Hao. 2025c. Llm-enhanced query generation and retrieval preservation for task-oriented dialogue. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14307–14321.
- Jiale Chen, Shunhao Li, baoshuo Kan, FuLee Wang, and Tianyong Hao. 2024. Leveraging intent entity enhancement for task-oriented dialogue. In *IJCNN2024*. IEEE.
- Xiaolin Chen, Xuemeng Song, Liqiang Jing, Shuo Li, Linmei Hu, and Liqiang Nie. 2023. Multimodal dialog systems with dual knowledge-enhanced generative pretrained language model. *ACM Transactions on Information Systems*, 42(2):1–25.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Zeyuan Ding, Zhihao Yang, Ling Luo, Yuanyuan Sun, and Hongfei Lin. 2024. From retrieval to generation: A simple and unified generative model for end-to-end task-oriented dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17907–17914.
- Xuelian Dong and Jiale Chen. 2023. Pludg: enhancing task-oriented dialogue system with knowledge graph plug-in module. *PeerJ Computer Science*, 9:e1707.
- Xuelian Dong, Jiale Chen, Chaoyu Wei, Chehao Ke, and Tianyong Hao. 2025. Mind: A unified multi-modal context-knowledge model for task-oriented dialogue. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Xuelian Dong, Jiale Chen, Heng Weng, Zili Chen, Fu Lee Wang, and Tianyong Hao. 2024. A new multi-level knowledge retrieval model for task-oriented dialogue. In *International Conference on Neural Computing for Advanced Applications*, pages 46–60. Springer.
- Xuelian Dong, Jiale Chen, Wenxiu Xie, Fu Lee Wang, and Tianyong Hao. 2026. Horizon: A unified framework for phase-wise retrieval-generation optimization in task-oriented dialogue system. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16972–16976. IEEE.
- Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.
- Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363.
- Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2020. Multidm-gcn: Aspect-guided response generation in multi-domain multi-modal dialogue system using graph convolutional network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2318–2328.
- Kaizhi Gao, Tianyu Wang, Zhongjing Ma, and Suli Zou. 2024. Winnie: Task-oriented dialog system with structure-aware contrastive learning and enhanced policy planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18021–18029.
- Peiming Guo, Sinuo Liu, Yanzhao Zhang, Dingkun Long, Pengju Xie, Meishan Zhang, and Min Zhang. 2025. An end-to-end model for photo-sharing multi-modal dialogue generation. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–7. IEEE.
- Jianheng Huang, Ante Wang, Linfeng Gao, Linfeng Song, and Jinsong Su. 2024. Response enhanced semi-supervised dialogue query generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18307–18315.
- Young-Jin Kim, Min-Jun Kim, Kyunghwan An, Jinwoo Ahn, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Eun-Sol Kim. 2024. Structure-aware multimodal sequential learning for visual dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13193–13201.
- Fanheng Kong, Peidong Wang, Shi Feng, Daling Wang, and Yifei Zhang. 2024. Tiger: A unified generative model framework for multimodal dialogue response generation. In *Proceedings of the 2024 Joint*

- International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16135–16141.
- Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Zhiyuan Ma, Jianjun Li, Guohui Li, and Yongjing Cheng. 2022. Unitranser: A unified transformer semantic representation framework for multimodal task-oriented dialog system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 103–114.
- Wang Meng, Suzan J Obaiys, Yeliz Karaca, and Nur Amalina Binti Jamaludin. 2026. Human–computer dialogue systems-based image-text visualization, fusion and multi-intent models’ construction. *Fractals*, 34(04):2540136.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu. 2020. Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6344–6354.
- Libo Qin, Xiao Xu, Lehan Wang, Yue Zhang, and Wanxiang Che. 2023. Modularized pre-training for end-to-end task-oriented dialogue. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1601–1610.
- Dongning Rao, Jietao Zhuang, and Zhihua Jiang. 2026. Leveraging dynamic few-shot prompting and ensemble method for task-oriented dialogue with subjective knowledge. *Information Processing & Management*, 63(2):104317.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Md Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. Dialogk: Knowledge-structure aware task-oriented dialogue generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2557–2571.
- Khaled Saab, Chunjong Park, Tim Strother, Jan Freyberg, David GT Barrett, Yong Cheng, Wei-Hung Weng, David Stutz, Nenad Tomasev, Anil Palepu, et al. 2026. Advancing conversational diagnostic ai with multimodal reasoning. *Nature Medicine*, pages 1–11.
- Tianyuan Shi, Liangzhi Li, Zijian Lin, Tao Yang, Xiaojun Quan, and Qifan Wang. 2023. Dual-feedback knowledge retrieval for task-oriented dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6566–6580.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Xin Tian, Yingzhan Lin, Mengfei Song, Siqi Bao, Fan Wang, Huang He, Shuqi Sun, and Hua Wu. 2022. Q-to-d: A query-driven task-oriented dialogue system. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7260–7271.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Deeksha Varshney and Asif Ekbal Anushkha Singh. 2021. Knowledge grounded multimodal dialog generation in task-oriented settings. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 421–431.
- Fanqi Wan, Weizhou Shen, Ke Yang, Xiaojun Quan, and Wei Bi. 2023. Multi-grained knowledge retrieval for end-to-end task-oriented dialog. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11196–11210.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449.
- Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. In *International Conference on Learning Representations*.

- Changzhu Xiong, Yuexian Zou, and Tianyu Li. 2026. Disentangled signals, dynamic prompts: A meta-network framework for robust task-oriented dialogue. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1996–2000. IEEE.
- 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*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Shiquan Yang, Rui Zhang, Sarah M Erfani, and Jey Han Lau. 2021. Unimf: A unified framework to incorporate multimodal knowledge bases into end-to-end task-oriented dialogue systems. In *IJCAI*, pages 3978–3984.
- Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Kang Zhang, Yu-Jung Heo, Du-Seong Chang, and Chang D Yoo. 2024. Bi-mdrg: Bridging image history in multimodal dialogue response generation. In *European Conference on Computer Vision*, pages 378–396.
- Bo Zhang, Jian Wang, Hui Ma, Bo Xu, and Hongfei Lin. 2023. Zrigf: An innovative multimodal framework for zero-resource image-grounded dialogue generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5464–5473.
- Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. Vista: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200.

A Related work

Traditional TOD systems rely on textual input and typically adopt a modular architecture that separates knowledge retrieval from response generation (Xiong et al., 2026; Dong et al., 2024; Chen et al., 2025b; Wu et al., 2019; Qin et al., 2020, 2023; Tian et al., 2022; Rony et al., 2022; Dong et al., 2026). DialoKG (Rony et al., 2022) incorporates structural information from knowledge graphs into language models by leveraging structure-aware knowledge embeddings and graph-weighted attention masks, enabling knowledge-aware response generation. MAKER (Wan et al., 2023) addresses retrieval inefficiency in KBs through hierarchical retrieval and generator distillation. LQ-TOD (Chen et al., 2025c) introduces a large language model-based query generation approach, using few-shot learning to create pseudo-queries and a retrieval retention mechanism to enhance overall performance. MUER (Chen et al., 2025a) enhances knowledge retrieval in TOD through a multi-view retriever that jointly selects relevant knowledge graphs and prunes irrelevant relations. It further introduces a memory tracking mechanism to reuse recently accessed knowledge, improving retrieval consistency across dialogue turns. Although these systems are easier to train, they often suffer from insufficient coordination across modules and are primarily designed for text-only TOD scenarios. This not only constrains model performance but also limits the extension of TOD systems to more realistic multimodal settings.

To better reflect real-world multimodal interactions, a few studies have begun incorporating visual information into TOD systems (Chang and Ko, 2023; Zhou et al., 2024; Zhang et al., 2023; Kim et al., 2024; Yoon et al., 2024; Firdaus et al., 2020; Yang et al., 2021; Guo et al., 2025). MultiDM-GCN (Firdaus et al., 2020) employs graph convolutional networks to fuse multimodal inputs and generate responses. UniMF (Yang et al., 2021) integrates a multimodal KB using multi-granularity fusion and external retrieval modules to support cross-modal reasoning. MKGITN (Varshney and Singh, 2021) proposed a multimodal knowledge-enhanced incremental Transformer model that generates more accurate responses by integrating dialogue context, external knowledge, and visual features. DKMD (Chen et al., 2023) uses a dual knowledge selection module to separately retrieve text-based and image-based knowledge by decoupling textual and

Method	Naturalness	Correctness	Task Completion
RAG + MLLM (Qwen3-VL-8B) [†]	2.96	3.49	3.42
RAG + MLLM (Qwen3-VL-8B) [‡]	3.05	3.20	3.34
RAG + LM (FlanT5-Large)	3.81	3.47	3.64
RAG + LLM (Qwen3-8B)	3.11	3.06	3.10
INSIGHT (FlanT5-Large)	4.06	4.24	4.19
INSIGHT (Qwen3-8B)	3.94	4.13	4.07

Table 6: Human evaluation results show the average scores on the VISTA dataset. [†]: zero-shot settings. [‡]: Settings for LoRA fine-tuning using cross-entropy loss.

visual dialogue inputs. MIND (Dong et al., 2025) introduces a text-image dialogue dataset and decouples the optimization of retrieval and generation. F2RVLM (Bi et al., 2026) explores fine-grained fragment retrieval in long-form multimodal dialogues by locating query-relevant utterance-image fragments. It adopts a generation-based VLM trained with supervised fine-tuning and GRPO-based reinforcement learning to improve retrieval accuracy and fragment-level semantic coherence. However, these methods represent entity attributes with a single image and rely solely on text for intent understanding, thereby limiting their capacity to model fine-grained visual semantics and complex mappings.

To alleviate these limitations, we propose the INSIGHT framework. It employs progressive dynamic fusion and graph-structured entity modeling to align textual and visual information, and adopts an end-to-end training strategy to jointly optimize retrieval and generation. This enables the model to generate high-quality multimodal responses in image-driven multi-turn dialogues.

B Human Evaluation

We conducted an online survey with 50 student volunteers majoring in computer-related fields, the results of the human evaluation are summarized in Table 6. All participants were asked to assess the naturalness, correctness, and task completion of responses generated by different models on a 5-point scale, based on the provided knowledge base, dialogue context, and golden response. The results demonstrate that the two methods using the INSIGHT framework achieved higher scores across all three evaluation dimensions, with the variant based on FlanT5-Large outperforming the one built on Qwen3-8B. This indicates that INSIGHT is better at understanding image-guided dialogue context and knowledge base, enabling it to generate multimodal responses that are both natural and capable of accurately completing user-specified tasks.

Method	Entities	Entity Retrieval	Image Retrieval	Total
Hierarchical Retrieval	30	0.8	1.2	2.0
	500	4.2	1.4	5.6
	1000	7.9	1.2	9.1
Global Retrieval	30	0.8	3.7	4.5
	500	4.5	30.9	35.4
	1000	8.1	90.7	98.8

Table 7: Inference latency comparison between our hierarchical semantic-perceptual retrieval and global retrieval. All latency values are reported in seconds.

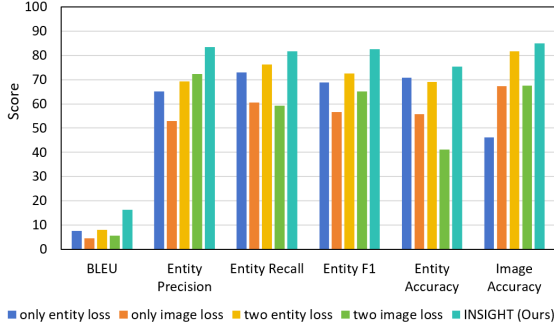


Figure 6: Performance comparison of different retrieval loss on the VISTA dataset.

C Inference Latency Analysis of Retrieval

To assess the practical efficiency of the proposed hierarchical semantic-perceptual retrieval mechanism, we compare its inference latency with a single-stage global retrieval baseline. The global retrieval baseline performs image retrieval over the entire image pool, whereas our hierarchical retrieval first localizes the most relevant entity and then conducts image retrieval only within the image set associated with that entity. As shown in Table 7, our hierarchical semantic-perceptual retrieval consistently achieves lower retrieval latency than global retrieval under different entity scales. Although it contains two retrieval stages, the entity-level retrieval reduces the candidate space for image retrieval, avoiding exhaustive matching over all images. These results demonstrate that our hierarchical semantic-perceptual retrieval is more scalable and suitable for practical deployment.

D Impact of Retrieval Loss

To evaluate the effectiveness of optimizing hierarchical semantic-perceptual retrieval under different loss functions, we conduct experiments using FlanT5-Large as the backbone, with results shown in Figure 6. The results demonstrate that INSIGHT achieves the best performance across all six metrics when employing our proposed strategy,

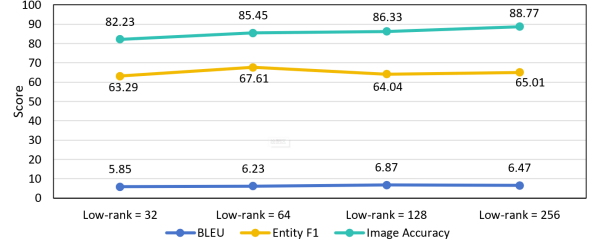


Figure 7: Performance of LoRA fine-tuning with different matrix ranks on the VISTA dataset.

namely optimizing the entity-level with $\mathcal{L}_{ent}$ and the image-level with $\mathcal{L}_{img}$. Moreover, neither using the same loss function for both hierarchical levels nor optimizing the two-level retrieval with a single loss yields optimal performance. These findings highlight the necessity and effectiveness of using distinct loss functions to separately optimize entity-level and image-level retrieval in hierarchical semantic-perceptual retrieval stage.

E Performance of LoRA Fine-Tuning with Different Matrix Ranks

As shown in Figure 7, experiments using the INSIGHT based on the Qwen3-8B with LoRA fine-tuning under different low-rank settings are conducted. The results demonstrate that varying the low-rank values leads to different impacts on performance. Specifically, when the low-rank=256, the model achieves the highest Image accuracy. For text response, the best Entity F1 is obtained with a low-rank of 64, while the highest BLEU is observed with a low-rank of 128. Therefore, in multimodal TOD, it is essential to explore and adjust the low-rank configuration based on the specific task requirements to achieve optimal performance.

F Case Study

Figure 8 presents case study on the VISTA dataset across different methods. First, when the user inquired whether the restaurant “*graffiti*” offered the dish shown in the image, both methods built upon our INSIGHT framework yielded affirmative responses, while the Qwen2.5-7B-based INSIGHT produced a more redundant reply. For the same question, under the zero-shot RAG + MLLM setting generated a lengthy response but failed to directly answer the user question. The RAG + LLM method produced a text response similar to the reference but incorrectly returned an unnecessary and irrelevant image response. The RAG + LLM

Does the graffiti serve the dish in the picture? It looks delicious! 		Yes, it is a popular specialty dish about the graffiti. Truth	
Of course, this is a specialty dish of the graffiti and is very popular with diners. INSIGHT (FlanT5-Large)	This is a specialty dish of the graffiti and is very popular with diners. I recommend you try it. Will there be anything else for you? It is delicious. I am going to recommend it to my friends. INSIGHT (Qwen2.5-7B)	The image you provided shows a plate of food that appears to be a traditional British breakfast, which includes sausages, baked beans, fried egg, toast, and black pudding. This type of meal is often associated with British cuisine. The restaurant "Graffiti" mentioned earlier serves British food. Zero Shot RAG + MLLM (Qwen2.5-VL-7B)	
Yes, it is a popular dish in the graffiti.  RAG + LM (FlanT5-Large)	Null RAG + LLM (Qwen2.5-7B)	Thank you for your visit. enjoy your meal. Fine Tuned MLLM (Qwen2.5-VL-7B)	
Is there availability of private rooms in the graffiti? User		Yes, the graffiti offers a private room in a unique ethnic style.  Truth	
Yes, the private room has a round dining table and a round chandelier.  INSIGHT (FlanT5-Large)	Yes, the private room has red chairs and walls and a bright chandelier over the dining table. You can choose to dine in privacy during your meal. Would you like any further information?  INSIGHT (Qwen2.5-7B)	Given your interest in British food, let's explore options in the west part of town. The Graffiti restaurant, which serves British food, is located at: Hotel Felix Whitehouse Lane Huntingdon Road. Zero Shot RAG + MLLM (Qwen2.5-VL-7B)	
Yes, the private room has a round dining table and a round chandelier. RAG + LM (FlanT5-Large)	Yes, it is a beautiful private room.  RAG + LLM (Qwen2.5-7B)	Yes, the room is very nice. Have a good day. Goodbye. Fine Tuned MLLM (Qwen2.5-VL-7B)	

Figure 8: A case study conducted on the VISTA dataset.

method did not provide an answer to the question. Meanwhile, the fine-tuned RAG + MLLM method generated a response unrelated to the user request. Furthermore, when the user asked whether “*graffiti*” had a “*private room*”, Only the two methods based on our INSIGHT provided correct text and image responses. In contrast, the other methods either failed to return the necessary image response or produced an incorrect one. Overall, these results demonstrate that our INSIGHT framework exhibits outstanding capability in addressing vision-guided TOD tasks.

G VISTA Dataset

G.1 The VISTA Dataset Examples

Figure 9 presents a multi-turn dialogue example from the VISTA dataset, along with the most relevant entity from the knowledge base. Specifically, when a user provides a top-down image of a dish and asks which korean restaurant serves it, the system must accurately interpret the visual semantics and retrieve multiple images of speciality dishes to identify those showing the same dish from different angles. Based on this, it should recommend the appropriate restaurant named cambridge lodge restaurant. Furthermore, when the user inquires

whether a given image of a attraction is located near the restaurant, the system needs to understand both visual and textual semantics to generate a natural and accurate text response and recommend the nearest relevant attraction as an image response.

G.2 Annotator Backgrounds

The VISTA dataset was annotated by 10 university students and 3 domain experts. The graduate students were all enrolled in computer science-related programs and had prior experience with data annotation tasks. They were responsible for constructing and annotating multimodal dialogues according to the predefined guidelines. The domain experts have research backgrounds in dialogue systems, multimodal learning, or natural language processing, and were primarily involved in designing the annotation guidelines, reviewing the annotated data, and conducting quality control. Their expertise ensured the consistency, accuracy, and reliability of the annotations.

G.3 Instructions Given to Participants

All participants involved in the annotation of the VISTA dataset were informed that their contributions were collected solely for academic research and dataset construction. They were instructed to

name: cambridge lodge restaurant, area: west, food: european, phone: 01223 355166, pricerange: expensive, location: 52.217736,0.102505, address: cambridge lodge hotel 139 huntingdon road city centre, type: restaurant, id: 19252, postcode: cb30dq, rooftop_bar: 	USER: Get me a restaurant that serves european food. SYSTEM: Which price range are you looking for? USER: Any price range is fine. It would be best if you could eat the dish in the picture. 
private_room: 	SYSTEM: Cambridge lodge restaurant serves european food and you can eat it. Is there anything else I can help you with? USER: Could I have the phone number and postcode for that restaurant please? SYSTEM: The phone number is 01223 355166 and the post code is cb30dq. Is there anything else? USER: Does the restaurant have a private room where we can eat? SYSTEM: Yes, cambridge lodge restaurant has a plush carpet and long tables available for private use. 
signature_food: 	USER: Whether or not you can drink at the restaurant? SYSTEM: Yes , cambridge lodge restaurant offers a rooftop bar for drinks. 
nearby_attractions: 	USER: Is this the closest attraction to the cambridge lodge restaurant? I want to go there. 
specialty_dishes:   	SYSTEM: No , the closest one is the one in the picture, which is a beautiful scenic attraction. 
	USER: Thank you. SYSTEM: You are welcome. Bye.

Figure 9: An example of a multi-turn dialogue and relevant entity from the VISTA dataset.

strictly follow the annotation guidelines and to ensure that all multimodal dialogues were grounded in the provided image-text knowledge base, without introducing unsupported entities, attributes, or external information. Moreover, participants were explicitly instructed not to include any personally identifiable or sensitive information. All images used during annotation were either sourced from publicly available resources or generated based on existing KB images, and no personal images were required or collected. The resulting dataset does not contain private or biometric data. All annotations were subject to manual verification and expert review to correct issues such as entity mismatches, ambiguous references, grammatical errors, and unclear semantic expressions, ensuring the quality and reliability of the VISTA dataset.

H Evaluation Metrics

BLEU (Papineni et al., 2002): it is a corpus-level statistical metric that measures the lexical similarity between the generated response and the human-annotated response. It calculates the n-gram overlap to assess the surface-level fluency. In the context of task-oriented dialogue, a higher BLEU score indicates that the system’s natural language genera-

tion aligns closely with human expressive patterns.

Entity Precision: it quantifies the accuracy of the content generated by the model. It is defined as the proportion of entities in the generated response that correctly match the entities present in the reference text. This metric is crucial for detecting hallucinations, ensuring the model does not provide incorrect information to the user.

Entity Recall: it measures the model’s ability to capture all necessary information required for the task. It represents the proportion of entities in the reference text that are recovered by the model. High recall ensures that the system provides a complete response without omitting critical constraints or service details requested by the user.

Entity F1 (Eric et al., 2017): it is the harmonic mean of Entity Precision and Entity Recall. Since there is often a trade-off between the precision and recall of entity generation, the Entity F1 serves as a robust, single-valued comprehensive metric to evaluate the model’s overall proficiency in identifying and generating target entities.

Entity Accuracy: it evaluates the performance of the entity retrieval. Specifically, it measures the ratio of cases where the entity retrieved from the knowledge base matches the ground-truth target

entity. This metric reflects the precision of the system’s internal reasoning and its ability to query the correct backend information.

Image Accuracy: it is a specialized metric for multi-modal dialogue systems. It denotes the proportion of dialogue turns where the image selected for the final response matches the target image in the reference. This metric evaluates the system’s cross-modal retrieval capabilities and the relevance of visual information within the dialogue context.

I Baselines

Qwen2.5-VL-7B (Wang et al., 2024): This vision-language model utilizes native dynamic resolution and multimodal rotary positional embeddings to capture features across various aspect ratios. By integrating visual tokens directly into a high-performance decoder, it excels in complex document parsing and long-video understanding.

Qwen3-VL-8B (Yang et al., 2025): A dense vision-language model that natively supports interleaved multimodal inputs of text, images, and video. It introduces enhanced interleaved-MRoPE for spatial-temporal modeling and DeepStack feature integration for tighter vision-language alignment, enabling strong multimodal reasoning and long-context understanding across document, video, and visual reasoning benchmarks.

Gemma3-4B (Team et al., 2025): A lightweight native multimodal model featuring an early-depth fusion architecture that replaces traditional projection layers for more efficient cross-modal interaction. It leverages knowledge distillation to achieve sophisticated reasoning capabilities within a compact 4B-parameter footprint.

Flan-T5 (Base & Large) (Chung et al., 2024): A encoder-decoder series optimized via extensive instruction tuning across 1,800+ tasks. It serves as a robust benchmark for sequence-to-sequence tasks, demonstrating exceptional stability in zero-shot generalization and linguistic comprehension.

Qwen2.5-7B (Yang et al., 2024): A decoder-only LLM pre-trained on an 18T-token corpus, specifically optimized for mathematics and programming. It supports a 128K context window through grouped-query attention, balancing extensive memory capacity with inference throughput.

LLaMA3-8B (Touvron et al., 2023): A milestone dense transformer model trained on 15T tokens, optimized for high-efficiency open-source deployment. Its implementation features a 128K

Tiktoken vocabulary and grouped-query attention, delivering state-of-the-art performance and scalability for 8B-parameter scale models.

Qwen3-8B (Yang et al., 2025): A dense decoder-only LLM in the Qwen3 series, designed to unify fast response generation and complex multi-step reasoning within a single framework. It supports dynamic switching between thinking and non-thinking modes, incorporates a configurable thinking budget mechanism, and extends multilingual capabilities to 119 languages and dialects, achieving strong performance on reasoning, coding, and agent-oriented benchmarks.