UniTranSeR: A Unified Transformer Semantic Representation Framework for Multimodal Task-Oriented Dialog Systems

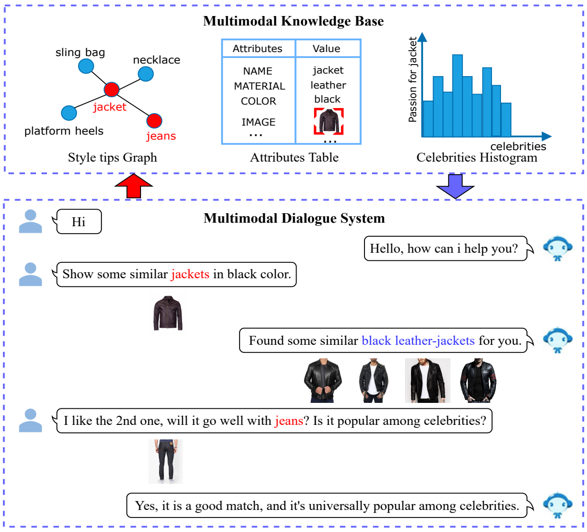
Zhiyuan Ma[1], **Jianjun Li**[1]*, **Guohui Li**[1], **Yongjing Cheng**[2]
[1] Huazhong University of Science and Technology (HUST), China
[2] National University of Defense Technology (NUDT), China
{zhiyuanma,jianjunli,guohuili}@hust.edu.cn
davidcheng1001@163.com

Abstract

As a more natural and intelligent interaction manner, multimodal task-oriented dialog system recently has received great attention and many remarkable progresses have been achieved. Nevertheless, almost all existing studies follow the pipeline to first learn intra-modal features separately and then conduct simple feature concatenation or attention-based feature fusion to generate responses, which hampers them from learning inter-modal interactions and conducting cross-modal feature alignment for generating more intention-aware responses. To address these issues, we propose UniTranSeR, a Unified Transformer Semantic Representation framework with feature alignment and intention reasoning for multimodal dialog systems. Specifically, we first embed the multimodal features into a unified Transformer semantic space to prompt inter-modal interactions, and then devise a feature alignment and intention reasoning (FAIR) layer to perform cross-modal entity alignment and fine-grained key-value reasoning, so as to effectively identify user's intention for generating more accurate responses. Experimental results verify the effectiveness of UniTranSeR, showing that it significantly outperforms state-of-the-art approaches on the representative MMD dataset.

1 Introduction

The multimodal task-oriented dialog systems are designed to help users achieve specific goals such as clothing recommendation or restaurant reservation, which is in growing demand in the current business environment. As a leading study, Saha et al. (2018) released a multimodal dialog dataset (MMD) in the online retail domain. Based on such a benchmark dataset, many multimodal dialog models incorporating domain knowledge have recently been proposed (Chauhan et al., 2019; Zhang et al., 2019, 2021), which basically exploit taxonomy-based method (Liao et al., 2018; Cui et al., 2019) or attention-based method (Nie et al., 2019; He et al., 2020) to incorporate knowledge base (KB) information for better performance.



Attributes	Value
NAME MATERIAL COLOR IMAGE ...	jacket leather black ...

Attributes Table

Hi

Hello, how can i help you?

Show some similar jackets in black color.

Found some similar black leather-jackets for you.

I like the 2nd one, will it go well with jeans? Is it popular among celebrities?

Yes, it is a good match, and it's universally popular among celebrities.

Figure 1: Example of multimodal task-oriented dialog including multimodal entity alignment and knowledge query from the MMD dataset (Saha et al., 2018). Note that red marks the entities to be queried in the multimodal knowledge base and blue marks the acquired knowledge information.

Though achieving remarkable progress, existing multimodal task-oriented dialog systems still suffer from the following three limitations. **Firstly**, prior models only learn the intra-modal features (including textual features, visual features and domain knowledge) separately before fusing them. Since these multimodal cues in general can enhance and complement each other, projecting them into a unified semantic space to learn the inter-modal features, with no doubt, can help improve the abilities of natural language understanding, which in turn will benefit the response generation. **Secondly**, prior models only conduct simple feature concatenation (Saha et al., 2018; Nie et al., 2019) or attention-based feature fusion (Cui et al., 2019) af-

[*] Corresponding author.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 103 - 114
May 22-27, 2022 ©2022 Association for Computational Linguistics

ter acquiring intra-modal representations, but without learning fine-grained alignment between different modalities before fusion, which is not favorable to query knowledge for accurate multimodal response generation. Take the dialog in Figure 1 as an example, when answering the user's query on similar style of jackets, the model is expected to align the word *"jackets"* with the corresponding visual features for proper semantic complement and entity enhancement. **Thirdly**, prior models basically lack the capability of entity-level reasoning, which prevents them from performing reasoning over crucial entities to guide intention-aware response generation. For example, in Figure 1, when the user asks *"show some similar jackets in black color"*, the chatbot is expected to properly explore the pivot attribute *"black"* that connects the start query cue *"jackets"* with the target recommended product images. Specifically, the model needs to perform a 2-hop reasoning over triples *(jacket_q, attribute, black_v)* and *(black_q, image, jacket_v)* and obtain the intended 4 images.

To address the aforementioned limitations, we propose a Unified Transformer Semantic Representation framework with feature alignment and intention reasoning, UniTranSeR for short. Specifically, to address the first limitation, we stand on the shoulder of Vision-and-Language Pre-training (VLP) methods (Lu et al., 2019; Li et al., 2019; Chen et al., 2020; Li et al., 2021) to propose a unified-modal Transformer encoder, which is used to project all the multimodal features into a unified semantic space to prompt inter-modality interactions, with the objective of learning better representations. Based on the unified encoder, we further address the second limitation by designing a feature alignment module to perform cross-modal feature alignment. Finally, to address the third limitation, we devise a fine-grained intention reasoning module for capturing users' real intentions, by leveraging a key-value attention based memory mechanism to perform multi-hop knowledge query for generating text or image responses.

We conduct experiments on MMD, one of the most influential benchmark datasets for multimodal dialog generation. We follow the mainstream evaluation script of dialog generation and demonstrate that UniTranSeR significantly outperforms the current state-of-the-art baselines. Ablation study also shows the efficacy of each component in improving the performance of dialog generation, and a further case study reveals that our model can effectively perform fine-grained token-level feature alignment for multimodal dialog generation.

2 Related Work

2.1 Unimodal Dialog Systems

Recent years has witnessed the remarkable success in textual dialog systems, which can be roughly divided into two categories: open-domain conversations with casual chi-chat (Song et al., 2020; Gangal et al., 2021; Chan et al., 2021; Yang et al., 2021) and task-oriented dialog systems (Pei et al., 2021; Santra et al., 2021; Wang et al., 2021; Mi et al., 2021; Madotto et al., 2021; Gou et al., 2021; Raghu et al., 2021), which are designed to help users achieve specific goals. Early efforts mainly adopt a sequence-to-sequence (Seq2Seq) architecture, but cannot work well in KB retrieval and reasoning. To alleviate this problem, copy mechanism (Eric and Manning, 2017) have been adopted and many memory augmented Seq2Seq models have been proposed (Bordes et al., 2017; Wen et al., 2018; Madotto et al., 2018; Wu et al., 2019; Reddy et al., 2019; Qin et al., 2019; Wang et al., 2020; Qin et al., 2020), which achieve promising results.

2.2 Multimodal Dialog Systems

With the flourishing of social media platforms, massive amounts of multimedia data are generated daily, which poses great demand for multimodal dialog systems. However, due to the lack of large-scale multimodal dialog datasets, researches in this domain have been limited. To this end, Saha et al. (2018) provided a vertical retail domain dataset MMD to promote the research and proposed a multimodal hierarchical encoder-decoder model (MHRED) as a baseline. Based on MHRED, Liao et al. (2018) incorporated the style tips into a knowledge-aware multimodal dialog model (KMD). Cui et al. (2019) designed a user attention-guided multimodal dialog system (UMD) by additionally considering the hierarchical product taxonomy and user's attention to products. Chauhan et al. (2019) introduced an ordinal and attribute aware multimodal dialog system (OAM) by employing a novel position and attribute aware attention mechanism. Later, Nie et al. (2019) proposed a multimodal dialog system with adaptive decoders (MAGIC), which can incorporate different forms of domain knowledge to generate different kinds of responses. Recently, combining with

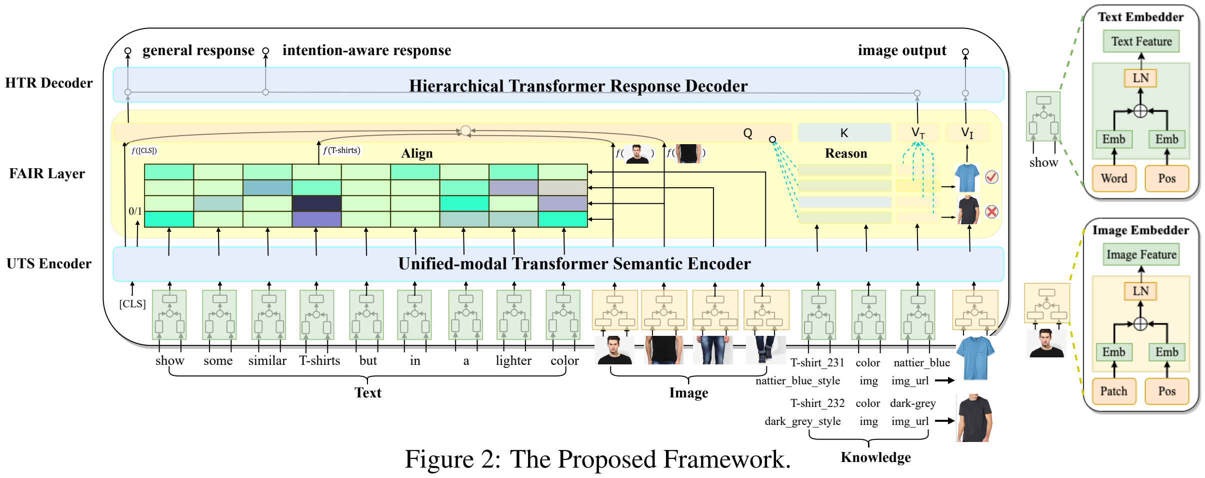


Figure 2: The Proposed Framework.

Transformer (Vaswani et al., 2017), He et al. (2020) advanced a multimodal dialog system via capturing context-aware dependencies of semantic elements (MATE) for textual response generation.

Most existing multimodal dialog systems learn intra-modal features separately for later feature concatenation or fusion. Different from them, our proposed UniTranSeR can project all the multimodal features into a unified semantic space to perform fine-grained feature alignment and intention reasoning, which can lead to more accurate responses. Vision-and-Language Pre-training (VLP) (Lu et al., 2019; Li et al., 2021) is another line of research relevant to our work, but different from ours in that it focuses more on boosting the performance of representation learning, while the multimodal dialog systems focus more on multi-turn multimodal interaction between users and agents.

3 Methodology

The proposed UniTranSeR mainly comprises three parts: Unified-modal Transformer Semantic (UTS) encoder (Sec. 3.1), Feature Alignment and Intention Reasoning (FAIR) layer (Sec. 3.2), and Hierarchical Transformer Response (HTR) decoder (Sec. 3.3), as shown in Figure 2. We define the multimodal dialog generation task as generating the most likely response sequence $Y = \{y_1, y_2, \cdots, y_n\}$ and selecting top-**k** most matched images, giving multimodal context utterances $U = \{u_1, u_2, \ldots, u_{|U|}\}$ and multimodal knowledge base B as inputs. The probability of a textual response can be formally defined as,

$$P(Y|U,B) = \prod_{t=1}^{n} P\left(y_t|y_1,\ldots,y_{t-1},U,B\right) \tag{1}$$

where y_t represents the current token decoded by the HTR decoder.

The UTS encoder is used to project all the multimodal features into a unified vector space for inter-modal interactions, while the FAIR layer is designed to align cross-modal hidden features, with textual features and visual features from previous UTS encoder as inputs. Similar to MAGIC (Nie et al., 2019), our HTR decoder is designed to decode three types of responses: general responses that refer to the highly frequent responses (e.g., courtesy greetings) in the conversation, such as *"How can I help you?"*; intention-aware responses that refer to the task-oriented utterances, such as *"Found some similar black leather-jackets for you"*; and multimodal responses that refer to the intention-aware responses with image output. The response type is determined by a query vector $\mathbf{Q}$ from the FAIR layer, in which an intention classifier is trained to decide which kind of response should be given out.

3.1 UTS Encoder

We first use a text embedder and an image embedder to extract textual features and visual features, respectively, and extract informative features from external knowledge by utilizing both text and image embedders. Afterwards, we feed these three kinds of features into a unified Transformer encoder for unified-modal semantic representation learning.

Text Embedder. To learn textual intra-modal features, we use a BERT tokenizer to split the input sentence into words and exploit a single transformer layer to obtain these words' initial embeddings. Note the self-attention mechanism in Transformer is order-less. So, it is necessary to encode the words' position as additional inputs. The final representation for each word is derived via summing up its word embedding and position embedding, followed by a layer normalization (LN) layer.

Image Embedder. To learn visual intra-modal features, we use a contour slicer to cut the input images into patches and exploit ResNet-50 (He et al., 2016) to extract these patches' visual features. We notice that people usually focus on four parts of a clothing image: head, upper body, lower body, and feet, so we intuitively use an equal-height mode to slice an image into four patches, which efficiently solves the problem of region feature extraction, without using complex target detection networks such as Faster R-CNN (Ren et al., 2015). Then, we feed the patches into ResNet-50 to get the patches' initial embeddings. Similarly, we also encode the position features for each patch via a 4-dimensional vector $[image\_index, patch\_index, width, height]$. Both visual and position features are then fed through a fully-connected (FC) layer, to be projected into the same embedding space. The final visual embedding for each patch is obtained by first summing up the two FC outputs, and then passing them through an LN layer.

Knowledge Embedder. To integrate informative features from external knowledge[1] into the task-oriented dialog, we equip the product knowledge base for each utterance through searching a fashion item table provided by MMD. We then treat these searched knowledge entries into the same triplet format, i.e., *(product, match, product)*, *(product, attribute, value)*, *(product, celebrity, passion_score)*. Next, for the text and image elements of these triples, we use the text and image embedders to obtain their respective representations.

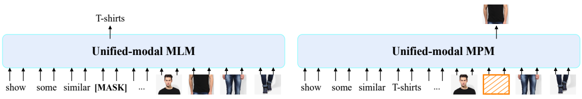


Figure 3: Illustration of MLM and MPM.

Unified Transformer Encoder. After obtaining the multimodal initial embeddings, denoted as $\mathbf{h}_t$, $\mathbf{h}_v$ and $\mathbf{h}_k$ respectively, we project them into a unified semantic space to obtain interactive representations by using a unified Transformer encoder. Specifically, in each utterance, the textual features, visual features and informative features correspond to l tokens with "[TXT]", 4 tokens[2] with "[IMG]" and 4 tokens[3] with "[KNG]". In order to integrate dialog history of previous rounds, we initialize the current [CLS]^p by using the representation of the previous round [CLS] $^{p-1}$. The output hidden state representations can then be phrased as:

$$\mathbf{H}^p = f\left([\mathrm{CLS}]^{p-1}\mathbf{h}_t^p[\mathrm{TXT}]\mathbf{h}_v^p[\mathrm{IMG}]\mathbf{h}_k^p[\mathrm{KNG}]\right) \tag{2}$$

where $f(\cdot)$ denotes the Transformer encoder, $\mathbf{H}_0^p$ denotes the hidden state representation of the current round [CLS]^p, which is regarded as the contextual semantic vector of the entire utterance in this round, $\mathbf{H}_{1:l}^p$ denotes the representations for the text sequence, $\mathbf{H}_{l+1:l+4}^p$ denotes the representations for the patch sequence, and $\mathbf{H}_{l+5:l+8}^p$ denotes the representations for knowledge entries. Note the superscript p is omitted for simplicity if no confusion occurs in the following discussion.

To obtain better representations, we introduce the Masked Language Modeling (MLM) loss and Masked Patch Modeling (MPM) loss to train them. We denote the input words as $w = \{w_1, \ldots, w_l\}$, the image patches as $v = \{v_1, \ldots, v_4\}$, the knowledge elements as $k = \{k_1, \ldots, k_4\}$, and the mask indices as $m \in \mathbb{N}^L$, where $\mathbb{N}$ is the natural numbers and L is the length of masked tokens. In MLM, we randomly mask out the input words with a probability of 15%, and replace the masked ones w_m with a special token "[MASK]", as illustrated in Figure 3. The goal is to predict these masked words by attentively integrating the information of their surrounding words $w_{\backslash m}$, image patches v and knowledge elements k, by minimizing the following loss:

$$\mathcal{L}_{\mathrm{MLM}}(\theta) = -\mathbb{E}_{(w,v,k)\sim U} \log P_\theta\left(w_m | w_{\backslash m}, v, k\right) \tag{3}$$

Similar to MLM, in MPM, we also randomly mask out the image patches and use zeros tensor to replace them, as shown in Figure 3. Unlike textual words that can be categorized as discrete labels, visual features are high-dimensional and continuous tensors, thus cannot be supervised via a negative log-likelihood loss. Following UNITER (Chen et al., 2020), we built the MPM loss as:

$$\mathcal{L}_{\mathrm{MPM}}(\theta) = \mathbb{E}_{(w,v,k)\sim U} g_\theta\left(v_m | v_{\backslash m}, w, k\right) \tag{4}$$

where v_m are masked image patches and $v_{\backslash m}$ are remaining patches. Note here g_θ is defined as an

[1] External knowledge of MMD includes: style tips graph, attributes table and celebrities histogram, as shown in Figure 1.

[2] Note when an utterance contains multiple images, it can be unrolled into a sequence of utterances, each containing a single image, the same as previous work.

[3] Including 3 textual features and 1 visual features.

L2 regression function, where

$$g_\theta\left(v_m | v_{\backslash m}, w, k\right) = \sum_{i=1}^{L} \left\| f_\theta\left(v_m^{(i)}\right) - \mathbf{h}_{v_m^{(i)}} \right\|_2^2 \tag{5}$$

3.2 The FAIR Layer

To align the cross-modal features for accurate intention classification and knowledge query, we devise a feature alignment and intention reasoning (FAIR) layer. In feature alignment, we use Image-Text Matching (ITM) and Word-Patch Alignment[4] (WPA) to conduct a two-level alignment. That is, ITM is used to align text and image in sentence-level, while WPA is used to align each split word and each sliced patch in token-level. In intention reasoning, we fuse $f([\text{CLS}])$ and aligned entities' hidden state representations to obtain a query vector $\mathbf{Q}$, which is then used for intention classification and knowledge query.

3.2.1 Feature Alignment

Image-Text Matching (ITM). In ITM, we use the output $f([\text{CLS}])$ of the unified Transformer encoder to compute the match probability of the sampled pair. Specifically, we feed $f([\text{CLS}])$ into an FC layer and a sigmoid function to predict a probability score $P_\theta(w, v)$, which is between 0 and 1. During training, we sample a positive or negative pair (w, v) from the dataset D at each step. The negative pair is created by randomly replacing the image or text in the same batch. We employ a binary cross-entropy loss for optimization:

$$\begin{aligned}\mathcal{L}_{\text{ITM}}(\theta) = -\,\mathbb{E}_{(w,v)\sim D}[y \log P_\theta(w, v) +\\ (1-y)\log\left(1 - P_\theta(w, v)\right)]\end{aligned} \tag{6}$$

where y is a binary truth label. Note here we only use ITM to train image-text pairs but without considering the knowledge vector, because it has already matched the textual sequence when being searched out.

Word-Patch Alignment (WPA). For more fine-grained alignment between each word and image patch, we introduce a WPA technology, which is used to train the consistency and exclusiveness between these cross-modal features to prompt alignment. We use a WPA loss to supervise the process, which is defined as:

$$\mathcal{L}_{\text{WPA}}(\theta) = -\sum\nolimits_{i=1}^{l} \sum\nolimits_{j=1}^{4} \mathbf{T}_{ij} \cdot \phi\left(w_i, v_j\right) \tag{7}$$

where ϕ denotes the $\cos(\cdot)$ similarity function, $\mathbf{T} \in \mathbb{R}^{l\times 4}$ is a ground truth table and each $\mathbf{T}_{ij} \in \mathbf{T}$ is a binary label 0 or 1. During training, we sample positive or negative pairs (w_i, v_j) from each multimodal utterance to construct a probability table, as shown in Figure 2. The above loss function $\mathcal{L}_{\text{WPA}}$ is then used to update the parameters θ. During inference, we continue to fuse aligned entities' hidden state representation and $f([\text{CLS}])$ to obtain a unified query vector $\mathbf{Q}$, which contains multimodal query information with entity enhancement, and will be used for subsequent intention reasoning.

3.2.2 Intention Reasoning

Intention Classify (IC). Given the query vector $\mathbf{Q}$, this component aims to understand the users' intention and thereafter determine which type of response should be generated. To be clear, there are a total of 17 types labeled in the MMD dataset, and each user's utterance is labeled with a specific intention type. Following MAGIC, we customize the type of response specifically for each intention, as shown in Table 1. Subsequently, we leverage an MLP layer to predict $\mathbf{Q}$'s probability distribution and select the highest probability to generate a response. Besides, a cross-entropy loss is applied to optimizing the intention classifier:

$$\mathcal{L}_{\text{IC}}(\theta) = \sum\nolimits_{i=1}^{|U|} \sum\nolimits_{j=1}^{17} I_{ij}^{*} \log P_\theta\left(I_{ij} \mid \mathbf{Q}\right) \tag{8}$$

where $P_\theta\left(I_{ij} \mid \mathbf{Q}\right)$ denotes the probability of being predicted as intention I_{ij}, and I_{ij}^{*} is a ground truth label. The intention classifier is trained by the loss function $\mathcal{L}_{\text{IC}}(\theta)$ to update parameter θ, and finally outputs a reliable intention prediction result I in the inference phase.

Knowledge Query (KQ). Given the predicted intention result I, this component first determines whether knowledge query is required based on Table 1. If required, we adopt a key-value memory mechanism to query all embedded knowledge triples[5]. Specifically, these embedded knowledge triples are divided into key parts and value parts, which are respectively denoted as vector $\mathbf{K}$ and vector $\mathbf{V}$. Note here $\mathbf{K}$ is obtained through a linear

[4] A modified version of the previous Word-Region Alignment (WRA), which can be adapted to the alignment between textual words and visual patches.

[5] The triple is in the form of $(head, relation, tail)$

Id	Intention categories	Response type	Component	Id	Intention categories	Response type	Component
1	greeting	general	IC	10	ask-attribute	intention-aware	IC+KQ
2	self-info	general	IC	11	suited-for	intention-aware	IC+KQ
3	give-criteria	multimodal	IC+KQ+MR	12	celebrity	intention-aware	IC+KQ
4	show-image	multimodal	IC+KQ+MR	13	filter-results	multimodal	IC+KQ+MR
5	give-description	multimodal	IC+KQ+MR	14	sort-results	multimodal	IC+KQ+MR
6	show-more	multimodal	IC+KQ+MR	15	switch-synset	general	IC
7	show-orientation	multimodal	IC+KQ+MR	16	buy	general	IC
8	show-similar	multimodal	IC+KQ+MR	17	exit	general	IC
9	goes-with	intention-aware	IC+KQ				

Table 1: The categories of user's intentions, their corresponding response types and required components.

fusion of the embedded head-entities and relations. The knowledge query process is as follows:

$$\alpha_i = \mathrm{Softmax}\left(\mathbf{Q}^{\mathrm{T}} \cdot \mathbf{K}_i\right) \tag{9}$$

$$\mathbf{V}_T = \sum\nolimits_{i=1}^{|M|} \alpha_i \mathbf{V}_i \tag{10}$$

where α_i denotes the attentive probability score for $\mathbf{K}_i$, $|M|$ is the number of knowledge triples, and $\mathbf{V}_T$ is a weighted sum of $\mathbf{V}_i$, which will be used for textual decoding in an intention-aware response.

Multi-hop Recommend (MR). Given the predicted intention result I and one-hop query result $\mathbf{V}_T$, this component first needs to determine whether an image recommendation is required based on Table 1. If required, we continue to use $\mathbf{V}_T$ as a query vector to perform another hop query over the entire knowledge base, which implies that the product images will be recommended, if the key parts of their corresponding triples have high similarity to $\mathbf{V}_T$. Specifically,

$$\beta_i = \mathrm{Softmax}\left(\mathbf{V}_T^{\mathrm{T}} \cdot \mathbf{K}_i\right) \tag{11}$$

After deriving β_i, we use $\mathbf{V}_I = \{q_i\}$, an image pointer vector, to select images with top β_i for recommendation, where

$$q_i = \begin{cases} 1, & \text{if } \mathbf{V}_i = \mathbf{1}_{1\times 512} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

and $\mathbf{1}_{1\times 512}$ is a column vector with each element equal to 1, which denotes for the special token [URL] of the image's link. Note here 512 is the embedding size in our unified Transformer encoder. It is not difficult to see that UniTranSeR can extend the above one-hop knowledge query to multi-hop by iteratively performing attention-based key-value reasoning and ultimately achieve multi-hop image recommendation.

3.3 HTR Decoder

As mentioned earlier, we used a hierarchy mechanism to decode different types of response sequences, including general responses, intention-aware responses and multimodal responses. They share the same uni-directional Transformer layer, but the semantic representations fed to this decoder are different. Specifically, for general responses, we just take the sentence-level representations f([CLS]) as input. For intention-aware responses, we take the concatenation of f([CLS]) and attentive vector $\mathbf{V}_T$ followed by an FC layer as input. For multimodal responses, we take the input for the intention-aware responses, as well as $\mathbf{V}_I$, the image pointer vector, as input.

Dataset Statistics	Train	Valid	Test
Dialogs	105,439	22,595	22,595
Proportion	70%	15%	15%

Table 2: Statistics of the MMD dataset.

4 Experimental Setup

4.1 Datasets and Metrics

To evaluate the performance of UniTranSeR, we conduct experiments on the widely-used benchmark dataset MMD contributed by Saha et al. (2018). The MMD dataset consists of over 150k conversations between users and chatbots in the retail domain, and each conversation describes a complete online shopping process. During the conversations, the user proposes his/her requirements in multimodal utterances and the chatbot introduces different products step by step until they make a deal. In our experiments, we follow Nie et al. (2019) to partition MMD. The statistics the dataset after partition are presented in Table 2, and more detailed statistics can be found in Appendix A.4.

Following several previous work (Nie et al., 2019; He et al., 2020; Zhang et al., 2021), we use Bleu-**n**, Nist and Recall@**k** to evaluate our model over two basic tasks separately, i.e., text task and image task. For the text task, we employ the proposed HTR decoder to produce all general responses and intention-aware responses. As the length of 20.07% target responses in MMD is less than 4, such as *"Hello!"* and *"Thanks a lot!"*, we follow Nie et al. (2019) to calculate Bleu-**n** by

Methods		Text Task					Image Task		
		Bleu-1	Bleu-2	Bleu-3	Bleu-4	Nist	Recall@1	Recall@2	Recall@3
Previous Methods	**MHRED** (Saha et al., 2018)	32.60	25.14	23.21	20.52	3.0901	0.7980	0.8859	0.9345
	KMD (Liao et al., 2018)	-	-	-	-	-	0.9198	0.9552	0.9755
	UMD (Cui et al., 2019)	42.78	33.69	28.06	23.73	-	0.9796	0.9980	0.9990
	OAM (Chauhan et al., 2019)	48.30	38.24	32.03	27.42	4.3236	-	-	-
	MAGIC (Nie et al., 2019)	50.71	39.57	33.15	28.57	4.2135	0.9813	0.9927	0.9965
	MATE (He et al., 2020)	56.55	47.89	42.48	38.06	-	-	-	-
Ours	**UniTranSeR**	**63.27**	**55.93**	**51.31**	**48.07**	**4.9774**	**0.9983**	**0.9995**	**0.9998**

Table 3: Main results. Relevance (higher better) between generated responses and golden responses. Note all our results are statistically significant with $p < 0.05$ under t-test.

varying $\mathbf{n}$ from 1 to 4. Note higher Bleu and Nist scores indicate that more $\mathbf{n}$-gram overlaps exist between the predicted and target responses, and hence are more favorable. For the image task, we adopt Recall@$\mathbf{k}$ to evaluate the efficacy of image response, where $\mathbf{k}$ is varied from 1 to 3. Note the image response is correct only if the positive image is recommended in the top-$\mathbf{k}$ product images.

4.2 Baselines

We compare our model with the following state-of-the-art baselines.

- **MHRED** (Saha et al., 2018)[6] is the first baseline work to integrate the visual features into a hierarchical encoder-decoder model for their constructed MMD dataset.
- **KMD** (Liao et al., 2018) incorporates the style tips into the memory augmented neural model and adopts deep reinforcement learning to boost the performance.
- **UMD** (Cui et al., 2019)[7] proposes a user attention-guided multimodal dialog system by considerring the hierarchical product taxonomy and the user's attention to products.
- **OAM** (Chauhan et al., 2019) proposes a novel ordinal and attribute aware attention mechanism for multimodal dialog generation.
- **MAGIC** (Nie et al., 2019)[8] adopts the adaptive decoders with intention understanding to explicitly generate three types of responses.
- **MATE** (He et al., 2020)[9] utilizes a multi-modal element-level encoder to integrate dialog context and leverages a knowledge-aware two-stage decoder for response generation, and achieves state-of-the-art performance.

4.3 Implementation Details

Following Saha et al. (2018) and Nie et al. (2019), we utilize two-turn utterances prior to the target response as the context, and set the vocabulary size to $26,422$. In our trainings, the batch size is set to 64, learning rate is set to $1e^{-4}$ and the max number of training epoches is set to $1e^{4}$. Adam optimizer is used to optimize all models. All experiments are conducted with PyTorch. More details about hyper-parameter settings can be found in Appendix A.1.

5 Evaluation Results

5.1 Response Quality Evaluation

Automatic Evaluation Following KMD, UMD and MAGIC, we evaluate model performance automatically from two aspects: text response and image response. From the results in Table 3, we can observe that our model UniTranSeR achieves the state-of-the-art performance on both tasks. Specifically, in text task, UniTranSeR exhibits the highest Bleu-$\mathbf{n}$ with varying $\mathbf{n}$ from 1 to 4 compared with other baselines, indicating that our model can generate responses closer to the golden ones. Moreover, our model outperforms MATE, a recent model that can capture context-aware dependencies of semantic elements, by 26.3% in Bleu-4 score, which verifies the effectiveness of our model in learning cross-modal feature alignment and conduct intention reasoning to generate more accurate and informative responses. In image task, an extremely difficult performance improvement can be observed, which further verifies the superiority of our model.

Human Evaluation The human evaluation mainly focuses on four aspects: fluency, relevance, correctness, and informativeness, which are all important for task-oriented dialogue systems (Cui et al., 2019; Nie et al., 2019; He et al., 2020). We first randomly selected 200 dialogs from the MMD datasets, and used different models to generate responses, including UMD, OAM, MAGIC, MATE

[6] https://github.com/amritasaha1812/MMD_Code

[7] https://github.com/ChenTsuei/UMD

[8] https://acmmultimedia.wixsite.com/magic.

[9] https://github.com/githwd2016/MATE/tree/dev

Model	Flue.	Rele.	Corr.	Info.	Overall Average	Achieve Ratio
UMD	2.25	2.84	3.20	2.20	2.62	54.1%
OAM	2.45	2.90	3.38	3.10	2.96	61.2%
MAGIC	2.20	3.15	3.45	3.88	3.17	65.5%
MATE	3.24	3.08	3.56	4.12	3.50	72.3%
UniTranSeR	**3.65**	**4.00**	**3.92**	**4.22**	**3.95**	**81.6%**
Golden	4.95	4.82	4.85	4.75	4.84	100%

Table 4: Human evaluation of responses on fluency (Flue.), relevance (Rele.), correctness (Corr.), informativeness (Info.) on randomly selected dialogs.

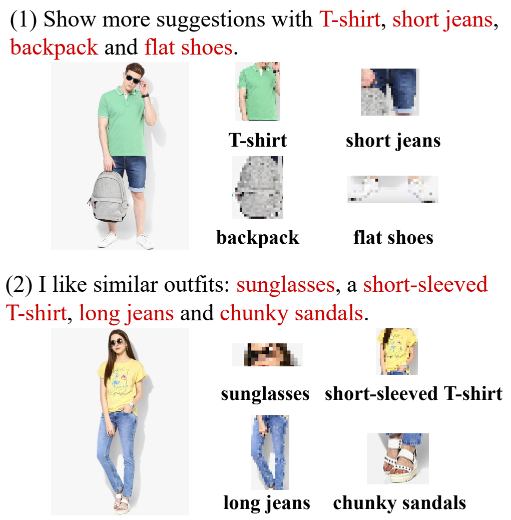
Methods		Bleu-4		Nist	
		Test	Δ	Test	Δ
UniTranSeR	**Complete**	**48.07**	-	**4.9774**	-
-UTS Encoder	-Trans.	42.07	12.48%	4.2620	14.37%
-HTR Decoder	-Trans.	45.35	5.66%	4.6291	7.00%
-FA Module	-ITM	40.20	16.37%	3.9580	20.48%
	-WPA	38.82	19.24%	3.5567	28.54%
-IR Module	-IC+KQ	21.65	54.96%	2.2804	54.18%

Table 5: Ablation study on MMD dataset.

and UniTranSeR. Then, we hired human experts to score the responses and golden responses in blind review on a scale from 1 to 5, which simulated a real-life multimodal task-oriented conversation scenario. By calculating the average score of the above metrics, we obtained the final manual evaluation results, as shown in Table 4. It can be observed that UniTranSeR consistently outperforms the other four models on all metrics, which is in line with the results of automatic evaluation.

5.2 Ablation Study

In this part, we perform ablation experiments to evaluate the effectiveness of each component. We focus on five crucial components and set them accordingly: 1) w/o UTS Encoder denotes that we use a BiGRU to replace the unified-modal Transformer encoder for multimodal encoding; 2) w/o HTR Decoder denotes that we use a Uni-directional GRU to replace the hierarchical Transformer decoder for response generation; 3) w/o ITM denotes that we remove the $\mathcal{L}_{\text{ITM}}$ loss to make the parameters not updated; 4) w/o WPA denotes that we remove the $\mathcal{L}_{\text{WPA}}$ loss and just regard the sentence-level representation $f([\text{CLS}])$ as query vector $\mathbf{Q}$ to query knowledge; 5) w/o IR Module denotes that we remove the IC and KQ components and just adopt the context vector $f([\text{CLS}])$ to generate responses[10]; From Table 5, we can observe that removing each component will result in a performance degradation. Specifically, w/o IR Module causes 54.96% drops in Bleu-4 score and 54.18% drops in Nist score, which verifies the great efficacy of intention classify and knowledge query components. Moreover, w/o WPA, w/o ITM and w/o UTS Encoder respectively cause 28.54%, 20.48% and 14.37% drops in Nist score, which further demonstrates the effectiveness of cross-modal feature alignment and unified-modal semantic encoding.


(1) Show more suggestions with T-shirt, short jeans, backpack and flat shoes.

(2) I like similar outfits: sunglasses, a short-sleeved T-shirt, long jeans and chunky sandals.

Figure 4: Visualization of Feature Alignment.

5.3 Case Study and Visualization

To better illustrate the advantage of our model and understand what the feature alignment module has learned, we visualize several examples of text-to-image attention, as shown in Figure 4. It can be observed that our model is able to capture fine-grained entity alignment between different modalities. The reason may be that: 1) We adopt a unified-modal Transformer semantic encoder, which enables to map different modalities of semantic cues into a same vector space to prompt inter-modality interactions for better representations; 2) Based on the obtained representations, the WPA technology can help supervise fine-grained word-patch alignment, which is beneficial to identifying user's real intention and generate more intention-aware responses.

6 Conclusion

In this paper, we propose a Unified Transformer Semantic Representation framework with feature alignment and intention reasoning, referred to UniTranSeR. Specifically, we project the multimodal features into a unified semantic space by utilizing a Transformer encoder to prompt inter-modal interactions. We further design a feature alignment and intention reasoning layer to conduct cross-modal feature alignment and fine-grained intention rea-

[10] Equivalent to generating general responses, since there is no knowledge query.

soning, with the objective of generating more accurate and intention-aware responses. Experiments on the representative MMD dataset demonstrate the effectiveness and superior performance of our UniTranSeR model in both automatic and human evaluation.

References

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Zhangming Chan, Lemao Liu, Juntao Li, Haisong Zhang, Dongyan Zhao, Shuming Shi, and Rui Yan. 2021. Enhancing the open-domain dialogue evaluation in latent space. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4889–4900. Association for Computational Linguistics.

Hardik Chauhan, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Ordinal and attribute aware response generation in a multimodal dialogue system. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5437–5447. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. 2019. User attention-guided multimodal dialog systems. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 445–454. ACM.

Mihail Eric and Christopher D. Manning. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 468–473. Association for Computational Linguistics.

Varun Gangal, Harsh Jhamtani, Eduard H. Hovy, and Taylor Berg-Kirkpatrick. 2021. Improving automated evaluation of open domain dialog via diverse reference augmentation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4079–4090. Association for Computational Linguistics.

Yanjie Gou, Yinjie Lei, Lingqiao Liu, Yong Dai, and Chunxu Shen. 2021. Contextualize knowledge bases with transformer for end-to-end task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4300–4310. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Weidong He, Zhi Li, Dongcai Lu, Enhong Chen, Tong Xu, Baoxing Huai, and Jing Yuan. 2020. Multimodal dialogue systems via capturing context-aware dependencies of semantic elements. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2755–2764. ACM.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2592–2607. Association for Computational Linguistics.

Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 801–809. ACM.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul A. Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7452–7467. Association for Computational Linguistics.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1468–1478. Association for Computational Linguistics.

Fei Mi, Wanhao Zhou, Lingjing Kong, Fengyu Cai, Minlie Huang, and Boi Faltings. 2021. Self-training improves pre-training for few-shot learning in task-oriented dialog systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1887–1898. Association for Computational Linguistics.

Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal dialog system: Generating responses via adaptive decoders. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 1098–1106. ACM.

Jiahuan Pei, Pengjie Ren, and Maarten de Rijke. 2021. A cooperative memory network for personalized task-oriented dialogue systems with incomplete user profiles. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1552–1561. ACM / IW3C2.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019. Entity-consistent end-to-end task-oriented dialogue system with KB retriever. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 133–142. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 6344–6354. Association for Computational Linguistics.

Dinesh Raghu, Atishya Jain, Mausam, and Sachindra Joshi. 2021. Constraint based knowledge base distillation in end-to-end task oriented dialogs. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 5051–5061. Association for Computational Linguistics.

Revanth Reddy, Danish Contractor, Dinesh Raghu, and Sachindra Joshi. 2019. Multi-level memory for task oriented dialogs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3744–3754. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 696–704. AAAI Press.

Bishal Santra, Potnuru Anusha, and Pawan Goyal. 2021. Hierarchical transformer for task oriented dialog systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5649–5658. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. 2020. Profile consistency identification for open-domain dialogue agents. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6651–6662. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Dual dynamic memory network for end-to-end multi-turn task-oriented dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4100–4110. International Committee on Computational Linguistics.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2021. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Haoyang Wen, Yijia Liu, Wanxiang Che, Libo Qin, and Ting Liu. 2018. Sequence-to-sequence learning for task-oriented dialogue with dialogue state representation. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3781–3792. Association for Computational Linguistics.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ze Yang, Wei Wu, Huang Hu, Can Xu, Wei Wang, and Zhoujun Li. 2021. Open domain dialogue generation with latent images. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14239–14247. AAAI Press.

Haoyu Zhang, Meng Liu, Zan Gao, Xiaoqiang Lei, Yinglong Wang, and Liqiang Nie. 2021. Multimodal dialog system: Relational graph-based context-aware question understanding. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 695–703. ACM.

Zheng Zhang, Lizi Liao, Minlie Huang, Xiaoyan Zhu, and Tat-Seng Chua. 2019. Neural multimodal belief tracker with adaptive attention for dialogue systems. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2401–2412. ACM.

A Appendices

A.1 Hyperparameters Setting

The hyperparameters used for MMD dataset are shown in Table 6.

Hyperparameter Name	MMD
Batch Size	64
Epoches	10,000
Text Embedding Size	512
Image Embedding Size	512
Transformer Embedding Size	512
Learning Rate	0.0001
Dropout Ratio	0.15
Teacher Forcing Ratio	0.9
Mask Length	6
Mask Probability	0.15
Replace Probability	0.15
Vocabulary Size	26,422

Table 6: Hyperparameters we used for MMD.

A.2 Description of Special Tokens

The special tokens used in our experiments are shown in Table 7.

A.3 Loss Function

Our total loss function $\mathcal{L}_{Total}$ comprises three parts: UTS encoder loss $\mathcal{L}_E$, FAIR layer loss $\mathcal{L}_F$ and HTR decoder loss $\mathcal{L}_D$, which can be calculated as follows:

$$\mathcal{L}_{Total} = \gamma_E \mathcal{L}_E + \gamma_F \mathcal{L}_F + \gamma_D \mathcal{L}_D \tag{13}$$

where γ_E, γ_F and γ_D are hyperparameters, and are initialized equally, i.e., 0.33, 0.33 and 0.33. Then, we tune them on the verification set to obtain a better weight setting of 0.30, 0.35 and 0.35.

The UTS encoder loss $\mathcal{L}_E$ contains two parts: $\mathcal{L}_{\text{MLM}}$ and $\mathcal{L}_{\text{MPM}}$,

$$\mathcal{L}_E = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{MPM}} \tag{14}$$

the FAIR layer loss contains three parts: $\mathcal{L}_{\text{ITM}}$, $\mathcal{L}_{\text{WPA}}$ and $\mathcal{L}_{\text{IC}}$:

$$\mathcal{L}_F = \mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{WPA}} + \mathcal{L}_{\text{IC}} \tag{15}$$

and the HTR decoder loss is divided into two types: the textual decoding loss $\mathcal{L}_{\text{TXT}}$ for text task and image recommend loss $\mathcal{L}_{\text{IMG}}$ for image task, which is consistent with previous work (Nie et al., 2019).

$$\mathcal{L}_D = \mathcal{L}_{\text{TXT}} + \mathcal{L}_{\text{IMG}} \tag{16}$$

Token	Description
[CLS]	Utterances classfication token
[TXT]	Text token
[IMG]	Image token
[KNG]	Knowledge token
[MASK]	Mask token
[URL]	Image link token
[PAD]	Padding token
[UNK]	Unknown token

Table 7: Description of special tokens in our experiments.

Dataset Statistics	Train	Valid	Test
Dialogs	105,439	22,595	22,595
Proportion	70%	15%	15%
Questions	2M	446K	445K
Image Responses	904K	194K	193K
Text Responses	1.54M	331K	330K
Avg. Utterances	40	40	40
Avg. Pos. Images	4	4	4
Avg. Neg. Images	4	4	4
Avg. Words in Question	12	12	12
Avg. Words in Response	14	14	14

Table 8: Detailed statistics of the MMD dataset.

A.4 Dateset Statistics

A detailed statistics of the MMD dataset is presented in Table 8.

A.5 Error Analysis

To better understand the limitations of our model, we conduct an error analysis on UniTranSeR. We randomly select 100 responses generated by UniTranSeR that achieve low human evaluation scores in the test set of MMD. We report several reasons for the low scores, which can roughly be classified into four categories. (1) KB information in the generated responses is incorrect (38%), especially when the corresponding equipped knowledge base is large and complex. (2) The sentence structure of the generated responses is incorrect and there are serious grammatical and semantic errors (24%). (3) The model makes incomplete response when there are multiple intentions contained in users' utterances (21%). (4) The model selects incorrect product images since different products have similar attributes (17%).