Sample Efficiency Matters: Training Multimodal Conversational Recommendation Systems in a Low Resource Setting

Anonymous ACL submission

Abstract

Multi-modal conversational recommendation (multi-modal CRS) can potentially revolutionize how customers interact with e-commerce platforms. Yet conversational samples, as training data for such a system, are difficult to obtain in large quantities, particularly in new platforms. Motivated by this challenge, we aim to design innovative methods for training multi-modal CRS effectively even in a low resource setting. Specifically, assuming the availability of a small number of samples with dialog states, we devise an effective dialog state encoder to bridge the semantic gap between conversation and product representations for recommendation. To reduce the cost of dialog state annotation, a semi-supervised learning method is developed to effectively train the dialog state encoder with a small set of labeled conversations. In addition, we design a correlation regularisation that leverages knowledge in the multi-modal domain database to better align textual and visual modalities. Experiments on two datasets (SIMMC and MMD) demonstrate the effectiveness of our method. Particularly, with only 5% of the MMD training set, our method (namely SeMANTIC) obtains better NDCG scores than those of baseline models trained on the full MMD training dataset.

1 Introduction

Recently, there has been a growing interest in conversational recommendation systems (CRS). These systems bring together the user-friendly nature of conversational AI and the business potential of recommendation systems, potentially revolutionizing how customers engage with e-commerce platforms. Unfortunately, conventional text-based dialogue systems have inherent limitations in capturing user preferences. In many practical situations, a blend of textual and visual cues allows agents to recommend products that are better aligned with user interests (e.g., see Figure 1 for an example).

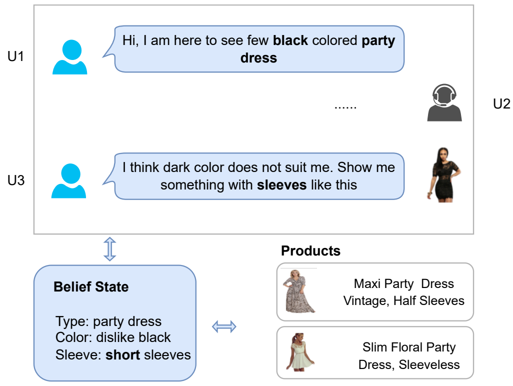

Figure 1: In a multimodal CRS, a user expresses her/his requirements with preferred example image. The dialog state (belief state) encapsulates user interest across turns and modalities.

The advance in deep learning along with the introduction of multi-modal benchmarks, such as MMD (Saha et al., 2018), have contributed significantly to the recent progress in multi-modal CRS. A number of methods have been developed using Recurrent Neural Networks (RNN) (Saha et al., 2018), RNN with attention (Cui et al., 2019), Graph Neural Networks (GNN) (Zhang et al., 2021), Memory Networks (Nie et al., 2021), Knowledge-enhanced Convolution Network (CNN) (Liao et al., 2018), and Transformer (Ma et al., 2022). Unfortunately, deep learning-based methods require a significant number of sample conversations with relevance annotation (for recommendation), which can be challenging to acquire. For example, the aforementioned methods have been trained on MMD using hundreds of thousands of conversations, and it is unclear whether these approaches remain effective when being trained with a smaller sample size.

In this paper, we examine multi-modal CRS in a low resource setting. Specifically, we consider that there is only a limited number of sample conversations and strive to make the most of the data by following two insights. Firstly, when the number of sample conversations is limited, augmenting them with dialog states can help align the representations of dialogues and products for better matching. This is supported by the fact that dialog state tracking (DST) is essential for traditional text-based task-oriented dialog (TOD) systems (Lei et al., 2018; Hosseini-Asl et al., 2020; Zhang et al., 2020; Yang et al., 2021). Unfortunately, dialog state annotation can be time-consuming, especially in multi-modal dialogs. Therefore, we assume that only a subset of sample conversations are annotated with dialog states, and design an effective method for dialog state modeling. Secondly, the vast amount of products with both textual and visual information should be exploited to bridge the cross-modal semantic gap. Intuitively, doing so helps improve the system's capability in understanding user preferences across modalities (see U3, Figure 1).

With such considerations, we propose a Sample Efficient Multi-modAl coNversaTIonal reCommendation system, or SeMANTIC for short. More specifically, dialog contexts and candidate products are first encoded with a context encoder and a product encoder separately, resulting in initial context/product representations. Such representations are then enhanced with Dialog-State Interaction modules that capture the interactions of the context (or the product) representations with shared dialog state embeddings. By doing so, we leverage dialog states to align the representations of the dialog and the product sides. Here, dialog state embeddings are learned via a teacher-student framework, where the teacher network has access to the limited size of dialogs with belief states, and the student network learns from the teacher to estimate dialog state embeddings from conversations without dialog states. We then propose a regularization term that makes state-aware (text/visual) representations of the same product closer to each other. By doing so, we effectively utilize the large number of products in the domain database for bridging the cross-modal semantic gap.

All in all, our main contributions are as follows:

- We propose a novel model, SeMANTIC, that enhances dialog and product representations with dialog states, and a regularization term that leverages the domain database to bridge cross-modal semantic gap.
- A semi-supervised learning is proposed based on the teacher-student framework to alleviate the dialog state annotation cost.
- Extensive evaluation on SIMMC and MMD datasets demonstrates the superiority of our model in comparison to strong baselines in a low resource setting.
- Further analysis validates that our semi-supervised learning approach is data efficient as it only requires a small ratio of supervision for learning dialog state embeddings.

2 RELATED WORK

2.1 MultiModal Conversational Systems

There have been a growing number of studies on multi-modal conversational systems thanks to the introduction of multi-modal datasets such as SURE (Long et al., 2023), FashionIQ (Wu et al., 2021; Yuan and Lam, 2021), MMD (Saha et al., 2018) and SIMMC (Kottur et al., 2021). Most of previous methods aim to enhance dialog representation using different network architectures (Saha et al., 2018; Ma et al., 2022; Nie et al., 2019; Zhang et al., 2021), external knowledge or side information (Cui et al., 2019; Nie et al., 2019; Zhang et al., 2021), mutual-information (Zhou et al., 2020), knowledge distillation (Jung et al., 2023), cross-modal interaction or attention (Cui et al., 2019; Ma et al., 2022).

Unlike these studies, we target an under-explored problem of learning effective representations with a limited number of conversations. It is noted that our focus is on grounding dialogs on external data (the recommendation task), which remains challenge particularly now that response generation can be greatly improved with large language models. As dialog systems are complicated, it is common for researchers to focus on substaks such as recommendation (Nie et al., 2021; Zhang et al., 2021), dense retrieval (Wu et al., 2023; Wang, 2024), Dialog State Tracking (DST) (Chen et al., 2020; Kumar et al., 2020) for deeper analysis.

2.2 Learning in a Low-Resource Setting

Deep learning has been the mainstream approach recently. Unfortunately, deep learning methods are also data hungry, requiring a large amount of training conversational samples with annotation. For example, to train a conversational recommendation system, it is needed to collect diverse dialog samples annotated with recommendations and various user requests (Budzianowski et al., 2018; Li et al., 2018; Liu et al., 2020). As labeled data is difficult to obtain, it is desirable to develop data efficient methods based on pretrained models (Yang et al., 2023; He et al., 2022), meta-learning (Dai et al., 2020), or semi-supervised learning (Yang et al., 2022; Huang et al., 2020; Li et al., 2020).

Our work falls into the semi-supervised learning category but focuses on multi-modal dialogs. To the best of our knowledge, our work is the first attempt at this important problem. It should be noted that we cannot simply adopt a unimodal method to a multi-modal scenario. For instance, one simple way to apply these available methods (Huang et al., 2020; Zhang et al., 2020) to our task is to consider DST as a text sequence generation task. However, as we empirically show in Section 5.3, without careful consideration of the semantic gap between modalities as well as between products and dialogs, even groundtruth (sequentialized) DST will not facilitate the recommendation task.

3 METHODOLOGY

Problem Formalization Let $\mathcal{D}_F$ be the set of M fully labeled dialogues $\tau_i = \{u_t | 1 \leq t \leq n_{\tau_i}\}$, where u_t indicates the t-th turn from either the user or the agent. Each (user or agent) utterance u_t contains the textual part u_t^T and the visual part u_t^I, i.e. a list of user uploaded images or system recommended product images. For t-th user turn, we are provided with a dialog state s_t^T that summarizes the user requests throughout the conversation. Additionally, let $\mathcal{D}_P$ be the set of partially labeled dialogs of which we do not have dialog state annotation. We assume that $\mathcal{D}_P$ is larger in size compared to $\mathcal{D}_F$, but still in a moderate size. The CRS task is formalized as selecting products from a domain database $\mathcal{P} = \{(\rho_k^T, \rho_k^I) | 1 \leq k \leq n_{\mathcal{P}}\}$ as response to a user request. Here, a product in $\mathcal{P}$ is associated with both textual description ρ_k^T and images ρ_k^I.

The overall architecture of SeMANTIC is depicted in Figure 2, where the main idea is to treat dialog states as shared (continuous) variables that bridge the semantic gaps between the textual modality and the visual modality, and between the conversation and the product sides. Specifically, representations of user texts/images and product texts/images are both enhanced with dialog state embeddings using Dialog State Interaction (DSI) modules (Section 3.2). Here, the dialog state embeddings are obtained by encoding the groundtruth dialog states for those in $\mathcal{D}_F$, and inferred by the dialog learner for those in the partially labeled set (Section 4). To mitigate the limited size of $\mathcal{D}_F$, we add a regularization term inferred from the partially labeled dialogs $\mathcal{D}_P$ and the abundance of products in the domain database $\mathcal{P}$ (section 3.4 and 4).

3.1 Context and Product Encoders

Context Encoder Let τ be a dialog context and $u_t^T = \{w_{t1}, w_{t2}, \ldots, w_{tn_t^T}\}$ be the textual utterance at the t-th turn, where w_{t_i} is an one-hot representation of the i-th word, we obtain the turn-level text representation as follows:

$$U_{ti}^T = w_{ti} W_{emb} + PE(i)$$

$$U_t^T = [U_{t1}^T, ..., U_{tn_t^T}^T]$$

$$\mathrm{v}_t^T = SumPool[SelfAttn(U_t^T, U_t^T, U_t^T)]$$

where W_{emb} is the word embeddings obtained from BERT (Devlin et al., 2018), PE and SelfAttn denote the position embedding and self-attention (Vaswani et al., 2017). The dialog-level representation for the textual modality is as follows:

$$V^T = [\mathrm{v}_1^T, ..., \mathrm{v}_{n_\tau}^T]$$

$$C^T = SelfAttn(V^T, V^T, V^T)$$

Similarly, we construct the turn-level visual representation from the t-th turn $u_t^I = \{I_{t1}, I_{t2}, \ldots, I_{tn_t^I}\}$:

$$U_{ti}^I = ResNet(I_{ti})$$

$$\mathrm{v}_t^I = SumPooling[U_{t1}^I, ..., U_{tn_t^I}^I]$$

$$V^I = [\mathrm{v}_1^I, ..., \mathrm{v}_{n_\tau}^I]$$

$$C^I = CrossAttn(C^T, V^I, V^I)$$

The final dialog-level representations c^T and c^I (for the textual and visual modalities) are attained from the last turn representations in C^T and C^I.

Product Encoder The product text ρ^T and visual ρ^I representations for a product $\rho_l = (\rho_l^T, \rho_l^I)$ are obtained similarly to the turn-level dialog representations (i.e. v_t^T and v_t^I). Note also that the low-level image representation ResNet are shared between the context encoder and the product encoder.

3.2 Dialogue State Interaction Module

Our objective is to exploit dialog states to align representations in multi-modal CRS. As such, we first get a dialog state embedding $S_0 \in R^{n_{state} \times n_{dim}}$ from the context (see Section 4 for more details). Inspired by Memory Networks (Sukhbaatar et al., 2015), we then introduce Dialog State Interaction (DSI) modules to enhance both dialog and product representations with information in dialog states.

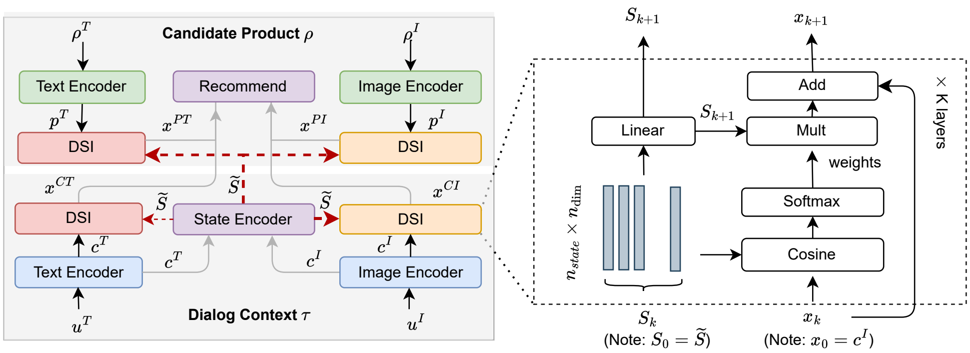

Figure 2: The overall architecture of SeMANTIC (left). Here, Dialog State Interaction (DSI) modules of the same color are shared between the dialog product sides. The details of a DSI module is shown on the right block.

The general architecture of Dialog State Interaction (DSI) module is depicted in Figure 2 with K layers of multi-hop interactions. Given an input vector x_k and a state embedding matrix S_k, the outputs of the k-th layer are obtained:

$$S_{k+1} = W_{k+1} S_k$$

$$a_{k+1,i} = \frac{cos(x_k, S_{k,i})}{\sum_j^{n_{state}} cos(x_k, S_{k,j})}$$

$$x_{k+1} = x_k + \sum_i^{n_{state}} a_{k+1,i} S_{k+1,i}$$

where W_{k+1} denotes the model parameters and a_{k+1} corresponds to the attention score vector. Note that x_0 is obtained from a context or product encoder (e.g. c^T, or p^T) and S_0 is from the state encoder module. As dialog state embeddings ($\tilde{S}$) are shared for the dialog context and the product candidate (see Figure 2), DSI module helps align the corresponding representations for effective matching.

3.3 Recommendation

Given a dialog τ and a candidate product ρ, the relevance score is measured as follows:

$$f(\tau, \rho) = \tanh[cos(x^{CT}, x^{PT}) + cos(x^{CI}, x^{PI})]$$

where $x^{CT}, x^{CI}, x^{PT}, x^{PI}$ are enhanced representations of the context and the candidate product, and extracted from the last layers of DSI modules.

3.4 Training

To train SeMANTIC, we construct a training set $\{(\tau_i, \rho_{ii}^+, \ldots, \rho_{in_{pos}}^+, \rho_{i1}^-, \ldots, \rho_{in_{neg}}^-)\}$ by sampling dialog contexts and the gold image responses from $\mathcal{D}_P$. Here, τ_i indicates one conversation context, whereas ρ_{ij}^+ and ρ_{ik}^- denote a positive recommendation and a (sample) negative recommendation for the i-th context. Note also that the dialog state encoder is trained jointly with the rest of the model. However, we postpone the detailed discussion until Section 4, where semi-supervised learning for dialog state modeling is described.

Ranking Loss The main objective for training SeMANTIC is to maximize the margin in the relevance score of the positive product compared to the negative product. In other words, we minimize the following rank loss:

$$\mathcal{L}_{rk} = max(0, 1 - f(\tau, \rho^+) + f(\tau, \rho^-))$$

where the loss is measured for a sample triple (τ, ρ^+, ρ^-). Here, we drop the context and product indices for simplicity.

Jensen Shannon Divergence To better align the context and the product representations, we measure Jensen-Shannon divergence (Menéndez et al., 1997) between the attention vectors extracted from the last layer of DSI (Equation 3.2 for $k = K$). Specifically, we respectively obtain (a^{CT}, a^{CI}) for the context text and images, and (a^{PT}, a^{PI}) for the product text and images, then measure:

$$g(\tau, \rho) = JS(a^{CT}, a^{PT}) + JS(a^{PI}, a^{PI})$$

Intuitively, we would like the g score to be small for the relevant pair (τ, ρ^+) and larger for the irrelevant pair (τ, ρ^-). To achieve this, we incorporate the following loss to the objective function:

$$\mathcal{L}_{JS} = max(0, g(\tau, \rho^+) - g(\tau, \rho^-))$$

Correlation Similarity Due to the limited size of conversational samples, we rely on the larger

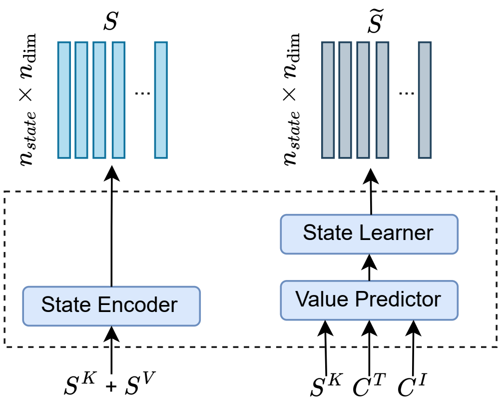

Figure 3: The State Encoder in the teacher SeMANTIC (left) vs that in the student SeMANTIC.

number of available products to bridge the gap between the textual and visual modalities. Our goal is to minimize the regularization term calculated for a given product ρ as follows:

$$\mathcal{L}_{co-sim}(\rho) = max(0, 1 - cos(x^{PT}, x^{PI}))$$

The idea here is make the (text/visual) state-enhanced representations of the same product closer to each other.

Overall Finally, the overall loss function $\mathcal{L}_{all}$ is:

$$\sum_i \left\{ \mathcal{L}_{rk} + \mathcal{L}_{JS} + \sum_{\rho_{ik}^{\pm}} \mathcal{L}_{co-sim}(\rho_{ik}^{\pm}) \right\}$$

where $\rho_{ik}^{\pm}$ indicates either a positive or negative sample associated with the context τ_i.

4 Semi-supervised State Learning

To leverage small samples with dialog states, we follow the teacher-student framework (Chen et al., 2017), where the teacher and student have a similar structure (Figure 2) but differ in the dialog state encoder (Figure 3).

Teacher State Encoder The teacher has access to the ground truth dialog state in $\mathcal{D}_F$, where each dialog state $u^S = [(u_i^{SK}, u_i^{SV})|1 \leq i \leq n_{state}]$ is a list of slot and value pairs. The slot keys are drawn from a predefined set of n_{state} product properties defined in the domain database $\mathcal{P}$, such as color or type. For each slot key such as color, the slot value is "none" if it is not mentioned in the dialog context τ_t, and a specific value (e.g. red) otherwise. For the i-th slot, we treat the slot key and value as strings and attain the key and value embeddings $S_i^K \in R^{1\times n_d}$, $S_i^V \in R^{1\times n_d}$ via BERT and MeanPooling, which is similar to the text encoder in Section 3.1. The state embedding is then obtained via self attention as follows:

$$S_i = S_i^K + S_i^V$$
$$S = [S_1, ..., S_{n_{state}}]$$
$$S = SelfAttn(S, S, S)$$

Student State Encoder The student network estimates the slot value embedding from the context information by employing a "Value Predictor". Specifically, we first obtain the key embedding $S^K \in R^{n_{state}\times n_d}$ for all slot keys similarly to that in the teacher state encoder. The value embedding are then calculated as follows:

$$\bar{C} = C^T + C^I$$
$$\widetilde{S}^V = CrossAttn(S^K, \bar{C}, \bar{C})$$

where CrossAttn is the cross attention operator. We then obtain the predicted state embedding $\widetilde{S}$ using the "State Learner" as follows:

$$\widetilde{S} = S^K + \widetilde{S}^V$$
$$\widetilde{S} = SelfAttn(\widetilde{S}, \widetilde{S}, \widetilde{S})$$

Joint Training We train the teacher network on $\mathcal{D}_F$ and the student network on $\mathcal{D}_F + \mathcal{D}_P$ using the loss function $\mathcal{L}_{all}$ as in Section 3.4. Hereafter, we refer to the teacher and the student training losses as $\mathcal{L}_{all}^{tea}$ and $\mathcal{L}_{all}^{stu}$. We then let the teacher guide the student network by minimizing the mean square error measured between groundtruth dialog state embeddings and the predicted state embeddings on $\mathcal{D}_F$. The joint training objective, therefore, is:

$$\alpha\mathcal{L}_{all}^{tea} + (1-\alpha)\left[\mathcal{L}_{all}^{stu} + \sum_{\tau_i \in \mathcal{D}_F} MSE(S_i, \widetilde{S}_i)\right]$$

where S_i, $\widetilde{S}_i$ are the outputs of the teacher and student encoders, respectively.

5 Experiments

Evaluation Datasets Experiments are conducted on MMD (Saha et al., 2018) and SIMMC (Kottur et al., 2021). The MMD dataset contains more than **150k** conversations in retail domain. Following previous works (Nie et al., 2021; Zhang et al., 2021), we adopt the updated MMD dataset constructed by Nie (Nie et al., 2021) and refer to it as MMD-v2, which is divided into training/validation/test sets with ratio 70%/15%/15%. To study the impact of the sample size and dialog states, we select around **7765** samples (5% of MMD-v2) and perform dialog state annotation with slot keys being product attributes. We refer to this set of MMD as MMD-v3. We split the data to sets train/valid/test so that the training/valid/test set of MMD-v3 is a subset of the corresponding set of MMD-v2. As for SIMMC, the dataset contains **10681** scene based conversations, which is divided into 68% for training, 16% for validation, and 16% for testing. We extend the multimodal coreference resolution task into a recommendation task by utilizing bounding boxes to extract product objects from the same scene.

Implementation Details We implement our proposed model using PyTorch[1] and conduct our experiments on 1 NVIDIA V100 GPU with a mini-batch size 64 and 50 epochs. The dimension of the initial word embedding is set to 768, and the dimension of the initial image embedding is set to 512. The dimensions of both context representation and product representation are set to 768. For each experimental setting, the results from multiple runs of SeMANTIC and the baselines are averaged.

Evaluation Metrics Following (Nie et al., 2021; Zhang et al., 2021), Precision@k, Recall@k, and NDCG@k for (k=5, 10, and 20) are the adopted metrics for the recommendation task in CRS.

Compared Methods We compare SeMANTIC to baselines with published codes including **MHRED** (Saha et al., 2018), **UMD** (Cui et al., 2019), **MAGIC** (Nie et al., 2019), **LARCH** (Nie et al., 2021), and **TREASURE** (Zhang et al., 2021). In addition, we also adapt **CLIP** (Radford et al., 2021), which is a popular image-text pretrained model, as one of our baseline. Details about the compared methods are given in the Appendix.

Experimental Design Our experiments are designed to answer the following research questions: 1) **RQ1**: How do SeMANTIC and other baselines perform when being trained with small conversational sample sets? (Section 5.1); 2) **RQ2**: How is the effectiveness of SeMANTIC when only smaller samples are labeled with dialog states? (Section 5.2); 3) **RQ3**: Do baselines effectively exploit dialog states if we provide them with grouthtruth dialog states during testing? (Section 5.3).

[1]https://pytorch.org/

5.1 Main Results

We consider the case when the number of conversational samples is in the scale of SIMMC or MMD-v3, which is much smaller compared to MMD-v2. Note that on MMD, all compared models are trained on MMD-v3 but tested on MMD-v3 or MMD-v2. In addition, we consider $\mathcal{D}_P = \mathcal{D}_F$ for SeMANTIC here, leaving the analysis for different ratios of these two sets to next section.

Table 1 presents the experimental results, where a number of observations can be drawn. Firstly, SeMANTIC outperforms the compared methods on SIMMC and two testing sets of MMD, partially validating its effectiveness and generalization. Secondly, while the unified memory network in LARCH may help bridge semantic gaps across modalities as well as between the conversation and product sides, the method may be too complex to be trained effectively with a small sample size. As a result, LARCH falls short compared to simpler methods like MHRED, MAGIC, and TREASURE, despite being the second best-performing method when being trained with the MMD-v2 training set (Nie et al., 2021). And finally, even though we train our method with MMD-v3, which is only 5% of the training set of TREASURE† (MMD-v2), the evaluation results on the test set of MMD-v2 show that our method is comparable to TREASURE† on NDCG@5, NDCG@10 , and even better on NDCG@20. It should be noted that training on MMD-v2 is time-consuming, thereby preventing us from training compared models multiple times for comparison. As a result, we directly report the results of TREASURE † from (Zhang et al., 2021).

Despite being a powerful pretrained model for image-text retrieval, CLIP does not perform well in our specific task and domain, particularly on MMD – the more challenging dataset compared to SIMMC. This highlights the importance of efficient methods for low-resource domain, of which data is not abundant for pretraining.

5.2 The Impacts of Sample Size

To verify the effectiveness of semi-supervised state learning, we conduct experiments on MMD-v3 and change the ratio of the sizes of $\mathcal{D}_F$ to $\mathcal{D}_P$. For every epoch, we first jointly train both teacher and student models on $\mathcal{D}_F$, then train the student model on $\mathcal{D}_P$ without considering ground-truth dialogue state. Figure 4 indicates that our model improves as more annotated data is utilized. Furthermore,

	Method	P@5	R@5	NDCG@5	P@10	R@10	NDCG@10	P@20	R@20	NDCG@20
						MMD				
MMD v3./ v3.	MHRED	34.56±1.50	40.91±1.83	39.09±1.35	20.54±0.79	48.55±1.92	42.60±1.33	12.14±0.42	57.35±1.94	45.82±1.31
	UMD	27.13±4.80	30.04±4.71	25.62±4.08	18.13±2.06	42.52±4.61	31.23±3.87	11.82±0.81	55.27±3.67	35.89±3.42
	MAGIC	46.33±0.77	53.48±0.94	51.61±1.87	26.21±0.34	60.72±0.83	54.86±1.55	14.39±0.19	66.93±0.93	57.10±1.44
	CLIP	14.10±0.19	16.96±0.33	16.81±0.37	8.71±0.12	20.88±0.43	18.63±0.41	5.47±0.08	26.11±0.52	20.60±0.43
	LARCH	30.64±2.57	37.00±2.93	36.66±3.25	21.22±1.23	50.23±2.77	43.56±2.94	13.01±0.36	61.25±1.59	48.00±2.53
	TREASURE	45.75±1.47	53.34±1.78	52.11±2.10	25.59±0.55	59.82±1.31	55.36±1.95	14.15±0.19	66.37±0.91	57.46±1.73
	SeMANTIC	**63.87**±0.39	**75.19**±0.54	**75.87**±0.71	**32.96**±0.16	**77.71**±0.53	**76.94**±0.72	**17.06**±0.09	**80.52**±0.47	**77.91**±0.71
MMD v3./ v2.	MHRED	30.66±3.00	35.30±3.71	36.47±3.31	18.51±1.43	44.08±3.36	39.87±3.22	10.97±0.64	52.29±3.08	42.85±3.09
	UMD	13.49±0.66	15.66±1.59	15.00±1.81	10.74±0.22	24.93±1.39	18.68±1.55	7.81±0.76	35.97±2.72	22.76±1.68
	MAGIC	38.31±1.77	44.88±2.06	43.38±2.60	22.08±0.62	51.86±1.44	46.46±2.34	12.48±0.22	58.85±1.02	48.96±2.16
	CLIP	12.08±0.32	14.82±0.29	15.39±0.33	7.22±0.19	17.64±0.31	14.37±4.89	4.49±0.11	21.81±0.37	18.24±0.37
	LARCH	23.61±1.42	28.55±1.66	29.39±1.95	16.90±0.52	40.02±1.16	35.32±1.71	10.71±0.12	50.41±0.56	39.51±1.44
	TREASURE	34.99±1.74	41.06±2.05	39.75±1.79	20.47±0.72	48.04±1.81	42.88±1.65	11.85±0.36	55.73±1.85	45.66±1.62
	SeMANTIC	**58.66**±0.32	**69.66**±0.34	**71.08**±0.65	**30.29**±0.09	**72.06**±0.17	**72.08**±0.59	**15.66**±0.06	**74.60**±0.24	**72.94**±0.59
	TREASURE †	59.87	71.39	71.24	31.34	74.85	72.72	16.33	78.17	72.87
						SIMMC				
	MHRED	22.93±0.51	67.20±1.41	51.16±1.30	14.46±0.22	85.83±1.12	57.14±1.18	8.27±0.04	94.57±0.45	60.24±1.01
	MAGIC	26.95±0.38	78.16±0.98	63.52±1.00	15.62±0.36	90.86±1.08	68.32±1.18	8.56±0.03	97.69±0.32	70.10±0.84
	CLIP	29.71±0.49	80.74±1.16	70.46±1.21	17.06±0.15	91.18±0.28	74.33±0.91	9.22±0.07	97.41±0.11	76.18±0.89
	LARCH	23.31±0.93	71.15±1.71	57.83±1.84	14.48±0.31	86.85±1.72	63.80±1.48	8.15±0.08	96.10±0.89	66.69±1.23
	TREASURE	27.50±0.47	79.43±1.00	64.99±1.31	16.00±0.18	91.66±0.57	69.89±1.24	8.60±0.04	98.10±0.16	71.27±1.07
	SeMANTIC	**31.99**±0.33	**87.14**±0.71	**76.82**±0.87	**17.85**±0.09	**95.45**±0.41	**79.96**±0.75	**9.35**±0.01	**98.99**±0.14	**81.04**±0.64

Table 1: The overall results of SeMANTIC and baselines, in which the average and standard deviations of different runs are reported. MMD v3/ v2 (or MMD v3/ v3) means we train the model on the training set of MMD-v3 and evaluate on the testing set of MMD-v2 (or MMD-v3). TREASURE† is both trained and tested on MMD-v2 and reported from (Zhang et al., 2021).

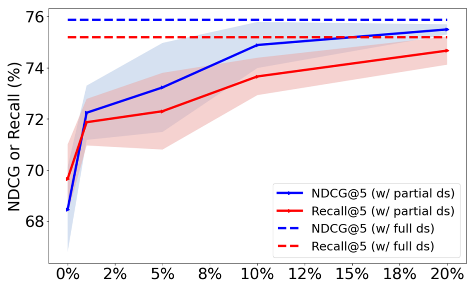

Figure 4: Performance of SeMANTIC trained with varying size of fully labeled data on MMD-v3.

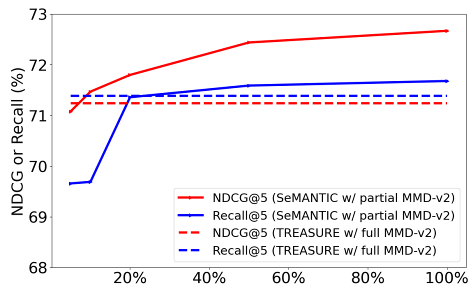

Figure 5: Performance of SeMANTIC trained with varying sample sizes on MMD-v2.

the reduction in standard deviation indicates that the model's performance becomes more stable as more samples with labeled states are considered. More importantly, our model's performance with 20% of the supervision ratio is nearly as good as having full supervision to learn state embeddings.

We evaluate the impact of the number of training (conversational) samples by conducting experiments on MMD-v2. Specifically, we keep $\mathcal{D}_F$ to be MMD-v3 training set, and increase the set $\mathcal{D}_P$ to include more samples from the training set of MMD-v2. The results of SeMANTIC and TREASURE are then reported on the testing set of MMD-v2 in Figure 5. The results show that SeMANTIC outperforms TREASURE in terms of NDCG@5 when the size of $\mathcal{D}_P$ to be around 10% of the MMD-v2, validating the sample efficiency of SeMANTIC.

5.3 Can Baselines Benefit from Dialog States?

We study whether the incorporation of dialog states into baselines can help improve performance of such methods. As adapting the baselines to incorporate dialog state prediction is nontrivial, we directly consider ground truth dialog states as part of the dialog input for the baselines during both training and testing. As SeMANTIC (w/ DS) only exploits groundtruth values during training, this setting gives baseline methods considerable advantage. This experiment is carried out on MMD-v3[2]. For SeMANTIC (w/o DS), state encoding excludes slot values during training, making it fair to compare with the baselines (w/o DS).

The performance comparison between the base-

[2]We skip the report on SIMMC due to similar observations

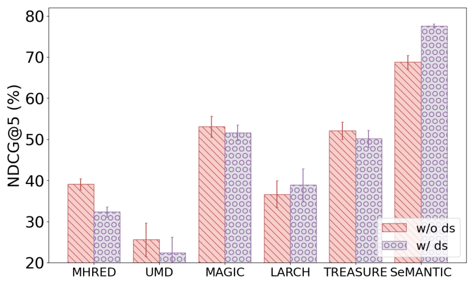

Figure 6: The impacts of dialog states on SeMANTIC and compared methods, tested on MMD-v3.

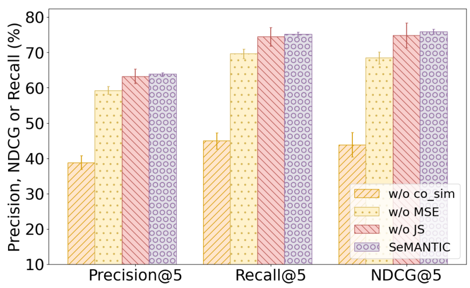

Figure 7: The impacts of different loss functions on SeMANTIC, tested on MMD-v3.

lines and SeMANTIC with and without dialog states is presented in Figure 6. Among all the methods, only LARCH and SeMANTIC show improvement on NDCG@k (k=5,10) when dialog states are considered. One possible explanation is that the slot values in dialogue states may not match product attribute values. As a result, only LARCH, which leverages diverse interactions between dialogs and knowledge, and SeMANTIC, which incorporates correlation similarity, can make good use of dialog state information.

5.4 Ablation Study

To examine the contributions of different loss functions, we exclude MSE loss (w/o MSE), correlation similarity loss (w/o co_sim), or JS divergence (w/o JS) from the training objective.

Figure 7 shows the impact of different loss types on SeMANTIC, measured on MMD-v3. The results reveal several findings. Firstly, the extraction of hidden information from text-image correlation in products (co_sim) and MSE loss are essential in enhancing the model's performance, given that the model's performance declines without this information. Secondly, the incorporation of $\mathcal{L}_{JS}$ helps reduce variation, making the model more stable. This is because excluding JS (w/o JS) leads to larger error bars in Figure 7.

Eval Metrics	Per Rec	Per Dialog
Win	32.20%	32.22%
Tie	63.84%	65.22%
Lose	5.98%%	2.56%

Table 2: Human evaluation for SeMANTIC vs TREASURE: the evaluation is measured per recommendation (per rec) or per dialog.

6 Human Evaluation and Case Study

To assess the effectiveness of our method, we conduct a human evaluation comparing its recommendation results against TREASURE (Zhang et al., 2021). We randomly sample 10 dialogues from MMD dataset, each has 6 recommendation turns on average. Three participants are then recruited, each is presented with recommendation results from both methods without revealing the method identities. We then count the ratio that SeMANTIC wins/ties/loses (to) TREASURE over all votes. The results of the human evaluation are summarized in Table2, demonstrating the superiority of our method over TREASURE. Please refer the Appendix for the case study.

7 CONCLUSION AND FUTURE WORK

This paper presents a novel approach named SeMANTIC for multimodal conversational recommendation systems (CRS). To align multi-modal representations, we propose dialog state interaction modules to enhance both the dialog and the product sides with dialog states. To overcome the challenge of collecting dialogue state labels, we develop a state value predictor to learn the dialog state embedding following a teacher-student framework. In addition, we introduce a correlation regularization for semantic alignment on the abundant products in the domain database. Our thorough experiments demonstrate the superiority of our proposed approach in the recommendation task when compared to existing methods.

Our method can be adapted to reduce the sample collection cost for general multimodal dialogues. For instance, one can consider dialog summaries instead of "dialog states" as the bridge for aligning multi-modal dialog representations. Those enhanced representations can then be used for downstream tasks such as external (textual/visual) knowledge retrieval or response generation.

Limitations

Due to time and computational constraints, our study did not carefully study the approach based on large vision-language models, such as (Radford et al., 2021; Li et al., 2023; Zhao et al., 2023; Wang et al., 2022). These models have shown promising results in various tasks, including semantic alignment and understanding in multimodal settings.

In the future, we plan to investigate how to efficiently and effectively adapt these large vision-language large models to our domain-specific database and explore their potential as base models for semantic alignment and recommendation in our multimodal conversational recommendation system. This would involve addressing challenges related to model scalability, computational resources, and efficient fine-tuning on domain-specific data.

By incorporating these advanced models, we aim to further enhance the performance and capabilities of our system, leveraging the rich information present in both textual and visual modalities.

Ethical Concerns

Our work is conducted using simulated data (published datasets), similar to previous studies (Saha et al., 2018; Cui et al., 2019; Nie et al., 2019; Zhang et al., 2021; Nie et al., 2021), and does not involve the use of any user-sensitive information.

During dialogue state annotation, we recruited participants from a crowd-sourcing platform and presented dialogue context, as illustrated in Figure1. Payment was adjusted appropriately considering the demographic profile of the participants. Additionally, we provided clear explanations regarding the utilization of the data.

The purpose of our research is to develop and evaluate a multimodal conversational recommendation system in a low resource setting. We recommend following data protection guidelines and regulations when applying our method in real platforms. It is crucial to obtain user agreements and informed consent before analyzing user requests or engaging in any data collection activities. This can be achieved through agree-upon interviews, and/or perform data simulation instead of using real conversations.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7521–7528.

Xiaolin Chen, Xuemeng Song, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. 2023. Dual semantic knowledge composed multimodal dialog systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1518–1527, New York, NY, USA.

Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. 2017. A teacher-student framework for zero-resource neural machine translation. *arXiv preprint arXiv:1705.00753*.

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*, pages 445–454.

Yinpei Dai, Hangyu Li, Chengguang Tang, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020. Learning low-resource end-to-end goal-oriented dialog for fast and reliable system deployment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 609–618, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Xinting Huang, Jianzhong Qi, Yu Sun, and Rui Zhang. 2020. Semi-supervised dialogue policy learning via stochastic reward estimation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 660–670, Online. Association for Computational Linguistics.

Yeongseo Jung, Eunseo Jung, and Lei Chen. 2023. Towards a unified conversational recommendation system: Multi-task learning via contextualized knowledge distillation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13625–13637.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

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

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33:8475–8485.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. *Advances in neural information processing systems*, 31.

Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable multimodal retrieval for fashion products. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1571–1579.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Yuxing Long, Binyuan Hui, Caixia Yuan, Fei Huang, Yongbin Li, and Xiaojie Wang. 2023. Multimodal recommendation dialog with subjective preference: A new challenge and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3515–3533.

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, Dublin, Ireland. Association for Computational Linguistics.

ML Menéndez, JA Pardo, L Pardo, and MC Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Liqiang Nie, Fangkai Jiao, Wenjie Wang, Yinglong Wang, and Qi Tian. 2021. Conversational image search. *IEEE Transactions on Image Processing*, 30:7732–7743.

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*, pages 1098–1106.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. *Advances in neural information processing systems*, 28.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Nguyen Wang, Zhang. 2024. Mitigating the impact of false negatives in dense retrieval with contrastive confidence regularization. In *Proceedings of the AAAI conference on artificial intelligence*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317.

Xing Wu, Guangyuan Ma, Meng Lin, Zijia Lin, Zhongyuan Wang, and Songlin Hu. 2023. Contextual masked auto-encoder for dense passage retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4738–4746.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. 2022. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238.

Yuting Yang, Wenqiang Lei, Pei Huang, Juan Cao, Jintao Li, and Tat-Seng Chua. 2023. A dual prompt learning framework for few-shot dialogue state tracking.

Yifei Yuan and Wai Lam. 2021. Conversational fashion image retrieval via multiturn natural language feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 839–848.

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 *Proceedings of the 29th ACM International Conference on Multimedia*, pages 695–703.

Yichi Zhang, Zhijian Ou, Min Hu, and Junlan Feng. 2020. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9207–9219.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1006–1014.

A Appendix

A.1 Dataset Statistics

In this paper, we conduct extensive experiments on two well-known datasets, namely MMD and SIMMC. For further insights, detailed statistics are provided in Table3 and Table4 respectively. Here, "Avg Rec Turns" indicates the average number of recommendations per dialog; and "Avg Pos Imgs" denotes the number of correct recommendations per turn whereas "Avg Neg Imgs" is the number of distractors for evaluation.

Dataset	**MMD v2**			**MMD v3 with DS**		
Dataset Stats	Train	Valid	Test	Train	Valid	Test
Dialogs	105439	22595	22595	5478	1113	1174
Proportion	70%	15%	15%	72%	14%	14%
Avg Rec Turns	5	5	5	6	6	6
Avg Pos Imgs	4	4	4	4	4	4
Avg Neg Imgs	616	618	994	628	632	989

Table 3: Statistics of the dataset by (Nie et al., 2019) (MMD v2) and the subset with dialogue state annotation (MMD v3 with DS).

Dataset	**SIMMC**		
Dataset Stats	Train	Valid	Test
Dialogs	7307	1687	1687
Proportion	68%	16%	16%
Avg Rec Turns	4	4	4
Avg Pos Imgs	2	2	2
Avg Neg Imgs	22	22	22

Table 4: Statistics of the SIMMC dataset.

A.2 Additional Experimental Results

Effect of Hyper-parameter α To study the effect of hyper-parameter α, we did several experiments with different α on MMD/ v3. The results with different α are given in Table5, which shows that our method is not sensitive to α.

Param α	**R@5**	**R@10**	**R@20**
α = 0.1	73.57±1.59	74.81±1.64	75.85±1.55
α = 0.3	74.04±1.64	75.27±1.69	76.22±1.67
α = 0.5	75.87±0.71	76.94±0.72	77.91±0.71
α = 0.7	75.65±1.71	76.77±1.79	77.74±1.73
α = 0.9	75.69±0.78	76.91±0.61	77.84±0.60

Table 5: The results with different α on MMD v3.

Varying Sizes of Conversational Samples In Section 5.2, to study the impacts of sample size, we show the performance of SeMANTIC trained with varying sample sizes on MMD-v2 in terms of NDCG@5 and Recall@5. Here, we further show the experiments in terms of NDCG@10 and Recall@10, and the results are provided in Figure9.

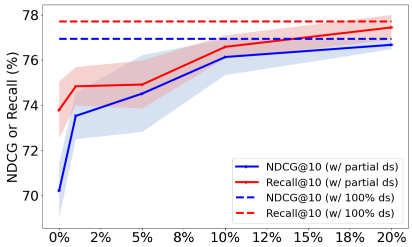

Figure 8: Performance in terms of NDCG@10 and Recall@10 for SeMANTIC trained with varying ratio of fully labeled data on MMD-v3.

Varying Size of Fully Labeled Data In Section 5.2, to study the impacts of sample size, we show the performance of SeMANTIC trained with varying ratio of fully labeled data on MMD-v3 in terms of NDCG@5 and Recall@5. Here, we further show the experiments in terms of NDCG@10 and Recall@10, and the results are provided in Figure8.

Furthermore, The results for changing the varying number of samples with dialog states (ds) on SIMMC dataset are presented in Table 6.

Ablation Study We further extend the ablation study to SIMMC dataset and Table 8 showcases more details of the impact of different loss types on SeMANTIC.

Human Evaluation and Case Studies To validate the effectiveness of our SeMANTIC, we presented a win case, a tie case, and a lose case in Figure 10. Additionally, we showcased the results of the TREASURE. Analysis of these retrieval results indicates our model's ability to accurately comprehend user intentions. Specifically, in Figure 10(a), SeMANTIC outperforms TREASURE by delivering the most correct images. Furthermore, in Figure 10(b), both SeMANTIC and TREASURE correctly select images, but SeMANTIC also arranges them at the top positions. In Figure 10(c), although SeMANTIC receives lower ratings in human evaluation, it consistently prioritizes global truth relevant items at the top positions. This observation underscores our model's proficiency in extracting pertinent information from utterance contexts to enhance understanding of user intentions for image response selection.

Figure 10: Top-10 image response selection results of our SeMANTIC and TREASURE in case win, tie and lose.

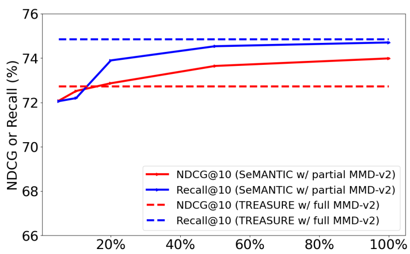

Figure 9: Performance in terms of NDCG@10, Recall@10 of SeMANTIC with different sizes of conversational samples.

A.3 Implementation Details

We implement our proposed model using PyTorch library [3] and conduct our experiments on 1 NVIDIA V100 GPU with a mini-batch size 64 and 50 epochs. Adam (Kingma and Ba, 2014) is adopted as the optimizer, with the initial learning rate 5×10^{-4} and the linear learning rate scheduler (Goyal et al., 2017) is used. Additionally, the dimension of the initial word embedding is set to 768, and the dimension of the initial image embed-

[3] https://pytorch.org/

	P@5	R@5	NDCG@5	P@10	R@10	NDCG@10	P@20	R@20	NDCG@20
SeMANTIC(0% labeled ds)	59.26±1.14	69.66±1.34	68.46±1.66	31.33±0.52	73.79±1.24	70.21±1.22	16.31±0.27	76.91±1.30	71.30±1.16
SeMANTIC(1% labeled ds)	61.08±0.72	71.87±0.91	72.23±1.06	31.76±0.37	74.83±0.85	73.52±1.03	16.47±0.19	77.69±0.98	74.52±1.04
SeMANTIC(5% labeled ds)	61.47±1.35	72.30±1.49	73.23±1.74	31.95±0.55	74.91±1.06	74.51±1.70	16.45±0.33	77.86±0.97	75.52±1.66
SeMANTIC(10% labeled ds)	62.56±0.56	73.66±0.73	74.89±0.90	32.48±0.19	76.59±0.51	76.13±0.80	16.89±0.07	79.75±0.42	77.20±0.77
SeMANTIC(20% labeled ds)	63.29±0.52	74.67±0.55	75.50±0.20	32.79±0.25	77.44±0.55	76.67±0.19	16.99±0.10	80.30±0.47	77.65±0.16
SeMANTIC(100% labeled ds)	63.80±0.39	75.19±0.54	75.87±0.71	32.96±0.16	77.71±0.53	76.94±0.72	17.06±0.09	80.52±0.47	77.91±0.71

Table 6: Performance of SeMANTIC on SIMMC when different size of labeled data is used for training.

	P@5	R@5	NDCG@5	P@10	R@10	NDCG@10	P@20	R@20	NDCG@20
MHRED(D_p 100%)	16.23	17.87	22.86	12.40	25.82	27.66	9.22	45.83	33.15
UMD(D_p 100%)	34.31	39.99	40.19	19.82	46.29	42.97	11.69	54.92	45.96
MAGIC(D_p 100%)	54.46	65.89	66.39	29.90	71.27	68.41	15.80	75.49	69.79
LARCH(D_p 100%)	55.01	65.82	68.29	29.99	71.61	71.21	15.95	76.20	73.02
TREASURE(D_p 100%)	59.87	71.39	71.24	31.34	**74.85**	72.72	**16.33**	**78.17**	72.87
SeMANTIC(D_F 5% and D_P 20%)	60.26	71.36	71.80	31.18	73.90	72.84	16.13	76.67	73.77
SeMANTIC(D_F 5% and D_P 100%)	**60.54**	**71.68**	**72.67**	**31.81**	74.71	**73.99**	16.24	77.62	**74.93**

Table 7: Detailed information about the performance of compared methods on MMD-v2, which are trained with different size of conversational samples for training.

ding is set to 512. The dimension of both context representation and product representation are set to 768. The number of layers of all transformer based encoders and decoders are set to 3, the number of attention heads in the multi-head attention is 8 and the inner-layer size is 768. We set all dropout rate to 0.1 (Srivastava et al., 2014), and α to 0.5 (Section 4). Moreover, we use 5 turns prior to the current turn as the context with the maximum sentence length of 30 and the maximum number of historical images to 5. It is worth mentioning that although both $\mathcal{L}_{all}^{teacher}$ and $\mathcal{L}_{all}^{student}$ contain $\mathcal{L}_{JS}$ and $\mathcal{L}_{co-sim}$, such losses are calculated by the teacher model and deactivated by the student model on $\mathcal{D}_F$. These losses are only activated for the student model on $\mathcal{D}_P$.

For CLIP, we only fine-tune its final linear projector and add self-attention layers to encoder turn level text embedding and image embedding. Then we concatenate text embedding and image embedding as the final context embedding and product embedding. For other baseline methods, we adhere to a standardized approach which adopts the default configurations as set in the original papers. By doing so, we ensure a consistent and accurate comparison with the established methodology.

A.4 Detailed Comparisons to Previous Methods

In the following, we provide detailed description on the compared baselines. In addition, we provide detailed discussion on previous methods that are closely related to our work but we are fail to conduct an empirical comparison as we do not have access to the original source code.

- **MHRED**: Saha et al. (2018) present a basic multimodal hierarchical encoder-decoder model (MHRED) as a first benchmark in the field of multimodal CRS.
- **UMD**: Cui et al. (2019) propose a user attention-guided multimodal CRS which is based on MHRED and uses a hierarchical product taxonomy tree to extract visual features.
- **MAGIC**: MAGIC (Nie et al., 2019) proposes knowledge-aware RNN to encode dialog context for response generation and product recommendation.
- **LARCH** Nie et al. (2021) introduce a contextual image search scheme (LARCH) with multi-form knowledge interactions via memory network.
- **TREASURE** Zhang et al. (2021) introduce TREASURE that represents dialog contexts using graph-based models and incorporate side information such as the product attributes and style-tips from celebrities.
- **UniTranSeR** (Ma et al., 2022) proposes a unified model based on Transformer to map image and textual modalities to a unified space.

				MMD					
Method	**P@5**	**R@5**	**NDCG@5**	**P@10**	**R@10**	**NDCG@10**	**P@20**	**R@20**	**NDCG@20**
w/o co_sim	38.84±1.98	45.02±2.29	43.90±3.51	21.87±0.92	50.84±2.21	46.52±3.21	12.11±0.44	56.47±2.11	48.55±3.04
w/o MSE	59.26±1.14	69.66±1.34	68.46±1.66	31.33±0.52	73.79±1.25	70.21±1.22	16.31±0.27	76.91±1.30	71.30±1.16
w/o JS	63.26±2.09	74.48±2.65	74.85±3.56	32.79±0.85	77.28±2.16	76.05±3.33	16.96±0.37	80.01±1.90	76.99±3.23
SeMANTIC	**63.87**±0.39	**75.19**±0.54	**75.87**±0.71	**32.96**±0.16	**77.71**±0.53	**76.94**±0.72	**17.06**±0.09	**80.52**±0.47	**77.91**±0.71
				SIMMC					
w/o co_sim	31.79±0.26	86.31±0.27	75.16±0.13	17.12±0.07	94.64±0.19	78.10±0.18	9.31±0.02	97.28±0.04	80.62±0.41
w/o MSE	31.03±0.19	86.44±0.36	75.23±0.48	17.19±0.02	94.74±0.13	78.00±0.42	9.31±0.01	97.18±0.11	80.73±0.39
w/o JS	31.27±0.37	87.01±0.80	76.74±1.15	17.21±0.10	95.38±0.46	79.34±0.99	9.34±0.01	98.33±0.06	81.09±0.88
SeMANTIC	**31.99**±0.33	**87.14**±0.71	**76.82**±0.87	**17.85**±0.09	**95.45**±0.41	**79.96**±0.75	**9.35**±0.01	**98.99**±0.14	**81.04**±0.64

Table 8: Effect of different loss functions.

As we fail to obtain their source code for empirical comparison, we analyze the method and find that this method is not designed for the multi-modal recommendation. Specifically, UniTranSeR first performs intention detection, then just uses the intent (textual modality) for product search. The experiments were conducted on MMD-v1 with much easier setting where the number of candidates is only 8 products.

- **MDS-S2** (Chen et al., 2023) recently introduced a novel method for multi-modal task-oriented dialog systems. The main idea is to exploit both the attribute and the relation information for external grounding knowledge retrieval, which is then used for text generation. The system is designed for external knowledge base that is more structured with well-defined attributes and relations. As both MMD and SIMMC do not fit this assumption, MDS-S2 has been tested on a newly constructed dataset.