PaCE: Unified Multi-modal Dialogue Pre-training with Progressive and Compositional Experts

Yunshui Li[1,2*†] **Binyuan Hui**[3*] **Zhichao Yin**[1,4] **Min Yang**[1‡] **Fei Huang**[3] **Yongbin Li**[3‡]

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]DAMO Academy, Alibaba Group
[4]University of Science and Technology of China
{ys.li, min.yang}@siat.ac.cn, {binyuan.hby, shuide.lyb}@alibaba-inc.com
http://github.com/AlibabaResearch/DAMO-ConvAI/pace

Abstract

Perceiving multi-modal information and fulfilling dialogues with humans is a long-term goal of artificial intelligence. Pre-training is commonly regarded as an effective approach for multi-modal dialogue. However, due to the limited availability of multi-modal dialogue data, there is still scarce research on multi-modal dialogue pre-training. Yet another intriguing challenge emerges from the encompassing nature of multi-modal dialogue, which involves various modalities and tasks. Moreover, new forms of tasks may arise at unpredictable points in the future. Hence, it is essential for designed multi-modal dialogue models to possess sufficient flexibility to adapt to such scenarios. This paper proposes **PaCE**, a unified, structured, compositional multi-modal dialogue pre-training framework. It utilizes a combination of several fundamental experts to accommodate multiple dialogue-related tasks and can be pre-trained using limited dialogue and extensive non-dialogue multi-modal data. Furthermore, we propose a progressive training method where old experts from the past can assist new experts, facilitating the expansion of their capabilities. Experimental results demonstrate that PaCE achieves state-of-the-art results on eight multi-modal dialog benchmarks.

1 Introduction

Enabling seamless communication between humans and machines is a long-standing goal of artificial intelligence research. The recent emergence of chatGPT [1] has increased confidence in the potential for achieving this goal. Beyond the use of textual language as a unique interface between humans and machines, perceiving and utilizing multi-modal information, especially visual information, has be-

---

*Equal contribution.
†Work done during an intern at Alibaba DAMO Academy.
‡Corresponding author.
[1]https://chat.openai.com/

Figure 1: An example of multi-modal dialogue, which involves multiple tasks, including multi-modal intent classification, multi-modal state tracking, multi-modal dialog retrieval and response generation.

come a crucial capability known as multi-modal dialogue (Shuster et al., 2020; Sun et al., 2021).

To facilitate the research on multi-modal dialogue, plenty of specific tasks and datasets have emerged in the community (Das et al., 2017; Shuster et al., 2018; Feng et al., 2022; Long et al., 2023). However, the overall quantity of data is still limited. Furthermore, multi-modal dialogue presents a greater challenge compared to traditional text-only dialogue track (Hui et al., 2021; He et al., 2022; Si et al., 2022), as it involves the integration of various modalities and more intricate task scenarios. As shown in Figure 1, the central tasks of multi-modal dialogue include multi-modal intent classification (Zang et al., 2021), multi-modal dialogue retrieval (Das et al., 2017; Zang et al., 2021),

13402

multi-modal dialogue state tracking (Liao et al., 2021), and multi-modal response generation (Kottur et al., 2021). Despite pre-training having become the consensus for multi-task learning in machine learning (Devlin et al., 2018; Radford et al., 2019, 2021), the research on pre-training models for multi-modal dialogue is an area that is yet to be fully explored.

In this paper, we focus on building pre-trained models of multi-modal dialogue. A key challenge is to unify different modalities and task forms, and make the best use of existing multi-modal dialog and non-dialog data. A recent popular trend on textual tasks is to build unified pre-trained foundation models by multi-task learning, e.g., T5 (Raffel et al., 2020). However, it attempts to mix all tasks learned from scratch thus is difficult to control the learning process, which is a completely black box. Although the Mixture-of-Experts (MoE) (Fedus et al., 2021; Du et al., 2022) architecture attempts to select independent experts for each input sample through token-level routing, it lacks specific semantics, i.e., it is entirely unknown what the experts are responsible for. We hope to find a new way to handle many multi-modal dialog tasks simultaneously and combine existing concrete skills to learn new tasks more efficiently.

To this end, we propose **PaCE**, a unified multi-modal dialogue pre-training framework with **P**rogressive **a**nd **C**ompositional **E**xperts. **First**, we decompose complicated multi-modal dialogue into fundamental sub-capabilities that could be learned with specific data. Different from traditional MoE, each expert in PaCE is tailored to one specific fundamental sub-capability of multi-modal dialogue, including CAPTION, CONTEXT, IMAGE, GROUNDING and GENERATION. **Second**, we propose a progressive pre-training strategy to evolve the model by controlling the combination of experts in different pre-training phases. Specifically, in stage I, we first train on multi-modal non-dialogue data to obtain CAPTION, IMAGE, and GROUNDING experts. In stage II, we train the CONTEXT expert, which is guided by the CAPTION expert on multi-modal dialog data to learn the dependencies in context. Furthermore, a dialogue GENERATION expert is derived by adding a response generation task based on the previously learned experts. **Third**, for pre-training PaCE, we collect a multi-modal dialog corpus with 1.4 million dialogs and a multi-modal non-dialog corpus with 4 million samples. Once
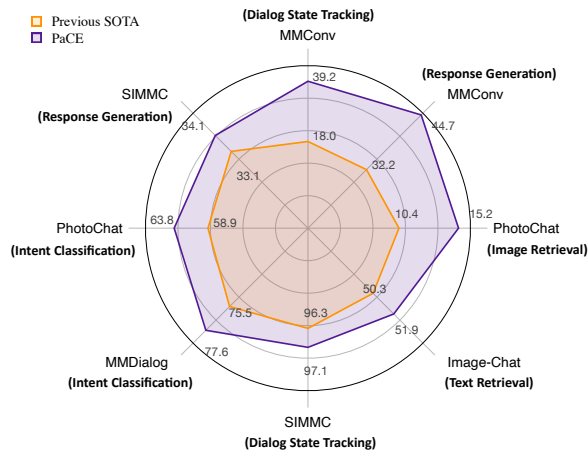

Figure 2: PaCE achieves state-of-the-art performances on a broad range of dialogue tasks compared with other customized or foundation models.

the pre-training of PaCE is finished, we can flexibly select different capability experts to solve a specific downstream task.

As illustrated in Figure 2, PaCE achieves state-of-the-art performance across a broad range of multi-modal dialogue benchmarks spanning four diverse downstream tasks, i.e., multi-modal intent classification, multi-modal dialogue retrieval, multi-modal state tracking, and multi-modal response generation This demonstrates that PaCE not only possesses a flexible model architecture but also exhibits adaptable training methodologies, resulting in remarkable performance.

2 Related Work

Pre-trained Vision-Language Models The pre-training paradigm, with its successes in natural language processing (Devlin et al., 2018; Radford et al., 2019), has sparked a revolution in Multi-modal Learning. ViLBERT (Lu et al., 2019) was the first work to adapt the BERT-like architecture for visual-language modeling, allowing for learning joint representation of images and texts. ViLT (Kim et al., 2021) constructed the vision module in the same way as the text module with a unified Transformer (Vaswani et al., 2017), eliminating the need for resource-intensive image feature extraction and significantly accelerating the model. CLIP (Radford et al., 2021) employed contrast learning to directly align images with natural language texts, eliminating the constraints of predefined image categories. ALIGN (Jia et al., 2021) and Florence (Yuan et al., 2021) further generalized this idea on noisier but larger image-text pairs. These models have demonstrated the ability to learn strong image and text representations for cross-

modal alignment tasks. In addition, a number of models (Cho et al., 2021; Wang et al., 2021, 2022; Yu et al., 2022; Alayrac et al., 2022) employed auto-regressive models to model the association between images and texts, using a unified generation approach to construct the task in an end-to-end manner. Although pre-trained vision-language models have shown promising results, they mainly focus on caption texts which are intrinsically different from human conversations (Kulhánek et al., 2021). To our best knowledge, the proposed PaCE model is the first multi-modal dialogue pre-training model.

Multi-Modal Dialogue Modeling   Numerous advanced works have been proposed along with the development of multi-modal dialogue datasets (Das et al., 2017; Mostafazadeh et al., 2017; Shuster et al., 2018; Zang et al., 2021; Zheng et al., 2021; Kottur et al., 2021; Liao et al., 2021; Feng et al., 2022). Several dialogue modeling works (Qi et al., 2020; Lee et al., 2021) have been conducted to improve the performance of conversational agents in image-grounded dialogue. Zang et al. (2021) proposed a dual-encoder model that utilized object labels to encode image features so as to perform a dialogue-based image retrieval task. Afterward, researchers (Yang et al., 2021; Chen et al., 2021) explored enriching textual expressions of generated dialogue responses through associative vision scenes. For textual response tasks, Zheng et al. (2021) proposed a multi-modal dialogue generation model based on Seq2Seq architecture, which was proved to be superior to the textual Seq2Seq model. Lee et al. (2022) proposed a joint multi-modal encoder-decoder model to incorporate visual inputs. However, the above models have demonstrated success in specific sub-tasks with a particular dataset, which cannot meet the requirements of a wide range of multi-modal dialogue tasks. To address this challenge, we propose a unified multi-modal dialogue pre-training model based on a divide-and-conquer strategy, which can combine different experts to complete a series of tasks.

3   Pre-training Data Construction

In this paper, we collect both multi-modal non-dialogue and multi-modal dialogue data for PaCE pre-training. The total statistics of our collected pre-training corpora are shown in Table 1.

Category	Dataset	Turns	Dialogs	Images
MultiNonDialog	CC3M	3.01M	-	3.01M
	SBU	867K	-	867K
	MSCOCO	113K	-	567K
	VG	108K	-	5.41M
MultiDialog	VisDial	1.2M	120K	120K
	Image-Chat	400K	202K	202K
	PhotoChat	97.6K	12.2K	11K
	MMConv	39.7K	5.1K	114K
	SIMMC2.0	117K	11K	1.5K
	MMDialog	4.82M	1.08M	1.53M

Table 1: Statistics of our collected pre-training corpora.

Multi-modal Non-dialogue Data (MultiNonDialog)   Similar to previous work (Kim et al., 2021), we first collect four multi-model non-dialogue datasets for image and text representation learning, including MSCOCO (Lin et al., 2014), VG (Krishna et al., 2017), SBU (Ordonez et al., 2011) and GCC (Sharma et al., 2018). In MultiNonDialog, each image is accompanied by one or more captions whose lengths are generally constrained to 20 tokens. Since GCC and SBU provide only image URLs, we collect the images via the given URLs which are still accessible.

Multi-modal Dialogue Data (MultiDialog)   We collect six existing multi-modal conversation corpora ranging from online forum chatting logs (Das et al., 2017; Shuster et al., 2018; Zang et al., 2021; Feng et al., 2022) to customer service conversations (Liao et al., 2021; Kottur et al., 2021) and build a large-scale multi-modal dialogue corpus. To ensure that each conversation has at least one corresponding image, we eliminate the text-only conversations from the original datasets. In addition, to satisfy the requirements of the Stage II pre-training, we use the BLIP model (Li et al., 2022b) implemented by Li et al. (2022a) to generate the appropriate textual caption for each image. The captions are constrained to 20 tokens.

4   Pre-training Method

Given a set of n multi-modal dialogue samples $\mathcal{D} = \{(U_i, R_i)\}_{i=1}^n$, where U_i and R_i represent the dialogue context and response, respectively. Compared to traditional textual dialogue, both $U_i = \{u_k^m\}_{k=1}^K$ and $R_i = \{r_q^m\}_{q=1}^Q$ can incorporate various types of information including textual texts and visual images, where K and Q are the number of elements, and $m \in \{t, v\}$ denotes the modality of U_i (or R_i). The notation t indicates textual utterances, while v indicates visual images.
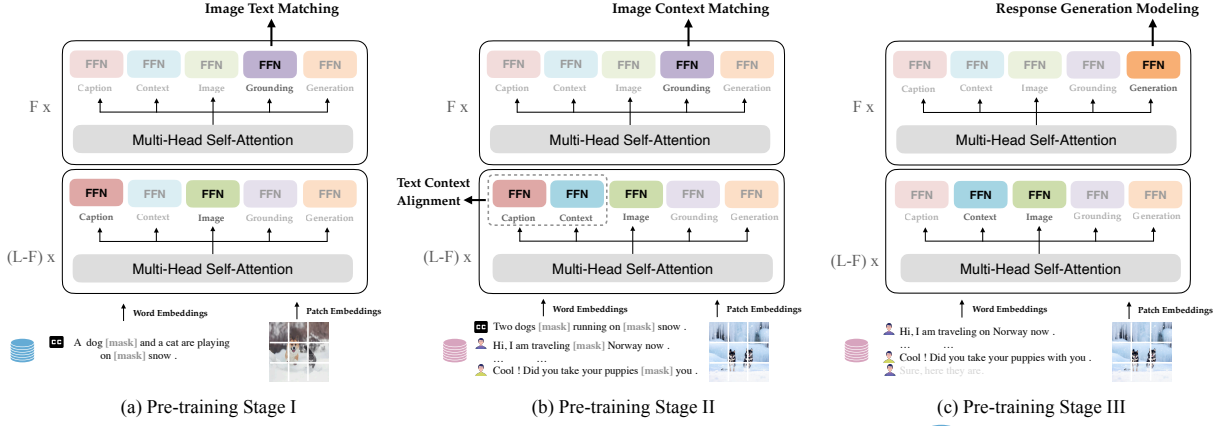
We devise a divide-and-conquer pre-training

Figure 3: Three-stage training based on different combinations of experts, where the 🗄 represents multi-modal non-dialog data and works mainly in the first stage, while the 🗄 represents multi-modal dialog data and works in the second and third stages. The 🆑 represents the caption of the input image.

strategy for multi-modal dialogue. Concretely, we decompose complicated multi-modal dialogue into five fundamental sub-capabilities and design five corresponding experts (i.e., CAPTION, CONTEXT, IMAGE, GROUNDING, and GENERATION experts). Then, we propose a progressive training strategy to evolve the model by controlling the combination of experts in different pre-training phases. Next, we describe the input representation learning module, the divide-and-conquer pre-training strategy, the pre-training objectives, and the fine-tuning process in detail.

4.1 Input Representation Learning

The proposed model is designed to handle input data from two modalities: visual representations and textual representations.

Visual Representations The dialogue context and response can be either visual or textual data. We use Vision Transformer (Dosovitskiy et al., 2020) to learn visual representations of images. Formally, we process the visual image $v \in \mathbb{R}^{H \times W \times \mathfrak{C}}$ by dividing it into $N = HW/P^2$ patches $v^p \in \mathbb{R}^{N \times (P^2 \mathfrak{C})}$, where $\mathfrak{C}$ is the number of channels, (H, W) is the resolution of the input image, and P is the patch resolution. This allows the model to extract meaningful features from the image by considering it as a set of small regions, rather than a single large array of pixels. The image patches are then flattened into vectors and processed by a linear projection using a weight matrix $\mathbf{W}_V \in \mathbb{R}^{(P^2 \cdot \mathfrak{C}) \times E}$ and a position embedding $\mathbf{W}_V^{\text{pos}} \in \mathbb{R}^{(N+1) \times E}$, resulting in patch embedding $\bar{v} \in \mathbb{R}^{N \times E}$, where E is the dimension of embedding. The position embedding is used to add additional information about the position of the

patch in the image. Finally, we obtain the visual representations $\boldsymbol{H}_0^v$ after summing patch embeddings and position embeddings.

Textual Representations The input text $t \in \mathbb{R}^{\mathfrak{L} \times |O|}$ is embedded into a dense representation $\bar{t} \in \mathbb{R}^{\mathfrak{L} \times E}$ by using a word embedding matrix $\mathbf{W}_T \in \mathbb{R}^{|O| \times E}$ and a position embedding matrix $\mathbf{W}_T^{\text{pos}} \in \mathbb{R}^{(\mathfrak{L}+1) \times E}$, where $|O|$ is the size of the vocabulary, $\mathfrak{L}$ is the length of text, and E is the dimension of embedding. It is noteworthy that we usually concatenate the context with the current utterance to form the final textual input. The textual representations can be denoted as $\boldsymbol{H}_0^t$.

4.2 Divide-and-Conquer Pre-training Strategy

We devise a novel pre-training strategy in a divide-and-conquer manner. Specifically, we first divide the complicated multi-model dialogue into several sub-problems, which can be learned in an easier way. The solutions to the sub-problems are then combined to give a solution to different downstream multi-modal dialogue tasks.

Multi-expert Architecture PaCE adopts an extension of the standard Transformer, which learns multiple semantic experts instead of a single feed-forward network (FFN) as in the original Transformer (Bao et al., 2021). Concretely, the experts share the information from both textual and visual modalities through a multi-head attention mechanism (MSA), while each expert FFN$^{\text{expert}}$ has its own unique parameters to learn a different semantic representation. Formally, the unique information, which is obtained by switching experts in each

block, can be formulated as:

$$\boldsymbol{H}'_l = \mathrm{MSA}\left(\mathrm{LN}\left(\boldsymbol{H}_{l-1}\right)\right) + \boldsymbol{H}_{l-1}$$
$$\boldsymbol{H}_l^{\mathrm{expert}_k} = \mathrm{FFN}^{\mathrm{expert}_k}\left(\mathrm{LN}\left(\boldsymbol{H}'_l\right)\right) + \boldsymbol{H}'_l \quad (1)$$

where $\boldsymbol{H}_{l-1}$ $(l \in [1, L])$ represents the output representation of the l-1 layer and L is the number of Transformer blocks. $\boldsymbol{H}_l^{\mathrm{expert}_k}$ is the representation of the k-th expert. The input representation could be formulated as $\boldsymbol{H}_0 = [\boldsymbol{H}_0^v, \boldsymbol{H}_0^t]$. Here, MSA and LN are the standard multi-head self-attention and layer normalization, respectively.[2]

Modality and Capability Experts   As illustrated in Figure 3, we divide the complicated multi-modal dialogue task into five easier sub-problems including CAPTION modeling, CONTEXT modeling, IMAGE modeling, GROUNDING, and GENERATION. We design a semantic expert to solve each sub-problem. These five experts can be divided into two categories: modality experts (CAPTION and IMAGE experts) and capability experts (GROUNDING, CONTEXT MODELING and GENERATION experts) tailored for multi-modal dialogue. Ultimately, we activate the modality and capability experts in a hierarchical manner, with the bottom $(L - F)$ layers activating only the modality experts and the top F layers activating the capability experts, where F is a pre-defined hyper-parameter.

Experts Combination for Different Tasks   We propose a progressive cascade pre-training strategy that solves different multi-modal dialogue tasks by adaptively combining the solutions to the sub-problems. We will introduce the details of progressive cascade pre-training in Section 4.3.

4.3   Pre-training Objectives

Our progressive cascade pre-training process consists of three phases, each with a tailored pre-training objective.

Stage I: Image-Text Matching   In stage I, similar to ViLT (Kim et al., 2021), we use non-dialogue multi-modal data D_n to learn the fundamental inter-modal alignment, and this stage involves only three experts, including the CAPTION expert, IMAGE expert and GROUNDING expert. As depicted in Figure 3(a), following word and patch embeddings, the text and image are separately processed into text and image representations by specialized CAPTION and IMAGE experts. These representations

---

[2] Due to limited space, we do not elaborate on MSA and LN, and readers can refer to (Vaswani et al., 2017) for implementation details.

are then fused and fed into the GROUNDING expert, yielding a unified representation of the image and text. We then employ the representation of the '[CLS]' token from the expert output as the input for a binary classification network to predict the alignment between the current text and image. The loss function of image-text matching is defined as:

$$\mathcal{L}_{\mathrm{itm}} = \mathbb{E}_{(V,T)\sim D_n}\mathrm{CE}\left(\boldsymbol{y}_{\mathrm{itm}}, \boldsymbol{p}_{\mathrm{itm}}(V, T)\right) \quad (2)$$

In addition to $\mathcal{L}_{\mathrm{itm}}$, we also employ the MLM loss $\mathcal{L}_{\mathrm{mlm}}$ in this stage for understanding unique textual modality. Concretely, following the method of BERT, we randomly select tokens in the text sequence and replace them with the [MASK] token. The model is trained to predict these masked tokens using the context of the remaining unmasked tokens and the visual clues. We adopt a masking probability of 15%. The final output vectors of the masked tokens are then fed into a classifier over the entire text vocabulary, with the training loss being the cross-entropy loss.

$$\mathcal{L}_{\mathrm{mlm}} = \mathbb{E}_{(V,\hat{T})\sim\{D_n\cup D_d\}}\mathrm{CE}(\boldsymbol{y}_{\mathrm{mask}}, \boldsymbol{p}_{\mathrm{mask}}(V, \hat{T})) \quad (3)$$

where $\hat{T}$ is a masked text, V is an original image and $\boldsymbol{p}_{\mathrm{mask}}(V, \hat{T})$ denotes the model's predicted probability for the masked token $\hat{T}$. D_n and D_d represent multi-modal non-dialogue and dialogue data, respectively.

The joint loss in stage I can be formulated as:

$$\mathcal{L}_{\mathrm{stage}}^{\mathrm{I}} = \mathcal{L}_{\mathrm{itm}} + \mathcal{L}_{\mathrm{mlm}} \quad (4)$$

Stage II: Image-Context Matching   In stage II, we use multi-modal dialogue data D_d to pre-train PaCE, which aims to model dialogue context for multi-modal dialogue tasks. At this stage, CAPTION expert will be activated in addition to the three experts from the first stage. Concretely, in the second stage, the dialogue context C is input to CONTEXT expert, the images V are input to IMAGE expert, and the corresponding image captions T are input to CAPTION expert. The loss function of image-context matching is defined as:

$$\mathcal{L}_{\mathrm{icm}} = \mathbb{E}_{(V,T,C)\sim D_d}\mathrm{CE}\left(\boldsymbol{y}_{\mathrm{icm}}, \boldsymbol{p}_{\mathrm{icm}}(V, T, C)\right) \quad (5)$$

In addition, we use the CAPTION expert learned in Stage I as a teacher to facilitate the learning of CONTEXT expert.

$$\mathcal{L}_{\mathrm{tca}} = \left\|\boldsymbol{H}_{L-F}^t - \boldsymbol{H}_{L-F}^c\right\|_2^2, \quad (6)$$

where $\boldsymbol{H}_{L-F}^t$ and $\boldsymbol{H}_{L-F}^c$ are the output of the $\{L-F\}$th-layer of CAPTION expert and CONTEXT expert, respectively.

Besides, we also employ MLM loss in stage II as defined in stage I, and the joint loss $\mathcal{L}_{\mathrm{stage}}^{\mathrm{II}}$ in

stage II could be formulated as:

$$\mathcal{L}_{\text{stage}}^{\text{II}} = \mathcal{L}_{\text{icm}} + \mathcal{L}_{\text{tca}} + \mathcal{L}_{\text{mlm}} \quad (7)$$

Stage III: Generation Modeling The third stage aims to enable the model to generate responses. The GENERATION expert is activated, and the input to this expert is composed of the CONTEXT expert and the IMAGE expert. The loss function in stage III is defined as follows:

$$\mathcal{L}_{\text{stage}}^{\text{III}} = -\sum_{n=1}^{N} \log \boldsymbol{p}_{\text{rgm}} \left(C_n \mid V, C_{<n} \right) \quad (8)$$

Here, we model generative capability by auto-regression, i.e., using past dialogue history $C_{<n}$ and associated images V to predict the current turn C_n of a dialogue.

4.4 Fine-Tuning on Downstream Tasks

Once the pre-training of PaCE is finished, we perform fine-tuning on specific downstream tasks. Thanks to our divide-and-conquer pre-training approach, we can flexibly select different capability experts to solve a specific downstream task. Specifically, for understanding tasks, including intent prediction, dialog retrieval, and dialog state tracking, we activate CONTEXT expert, IMAGE expert, and GROUNDING expert. For the generation task, i.e. dialog state tracking, and response generation, we activate the CONTEXT expert, IMAGE expert, and GENERATION expert.

5 Experiments

5.1 Downstream Datasets

To comprehensively evaluate our PaCE, we conduct extensive experiments on seven datasets belonging to four downstream tasks.

Multi-Modal Intent Prediction For multi-modal intent prediction, PhotoChat (Zang et al., 2021) and MMDialog (Feng et al., 2022) are selected as benchmark datasets. This task aims to identify the specific intent of the user in the multi-modal context. More specifically, it predicts the probability of photo sharing in the upcoming conversation turn.

Multi-Modal Dialog Retrieval For text-to-image retrieval, we select PhotoChat (Zang et al., 2021) as our benchmark dataset. It encompasses 12k dialogues, each accompanied by a user photo exchanged during the conversation. The goal of this task is to select the most appropriate photo given the dialog context. For image-to-text retrieval, we

select Image-Chat (Shuster et al., 2018) to evaluate our model, which consists of 202k dialogues over 202k images.

Multi-Modal Dialog State Tracking MM-Conv (Liao et al., 2021) and SIMMC2.0 (Kottur et al., 2021) datasets provide a good base for carrying out multi-modal dialog state tracking. The MMConv dataset contains 5.1k dialogues collected by enabling multi-modal conversations between human-to-human role-playing pairs under real-life traveling scenarios. In contrast, the SIMMC2.0 corpus includes 11,000 task-oriented dialogs in the shopping domain that are grounded in immersive and photo-realistic contexts.

Multi-Modal Response Generation Generating appropriate responses for satisfactory task completion is the ultimate goal of task-oriented dialogue agents. In this task, we selected MMConv (Liao et al., 2021) and SIMMC2.0 (Kottur et al., 2021) as our benchmark datasets.

5.2 Experimental Setting

We use the *bert-base-uncased* tokenizer to tokenize text inputs. We learn the textual embedding-related parameters from scratch, instead of fine-tuning them from pre-trained BERT. For all experiments, we use AdamW optimizer (Loshchilov and Hutter, 2017) with base learning rate of 10^{-4} and weight decay of 10^{-2}. The learning rate is warmed up for 10% of the total training steps and is decayed linearly to zero for the rest of the training. We set the total number of the Transformer layers L to 12, with the number of layers F for the top Transformer set to 3. We initialize the Transformer weights with the pre-trained ViT (Dosovitskiy et al., 2020). In the pre-training process, we utilize 200K steps, 25K steps, and 10K steps, respectively, for the three stages on 8 NVIDIA A100 GPUs with a batch size of 4,096.

5.3 Evaluation Metrics

For intent prediction, we adopt the F1 score as the evaluation metric to measure the effectiveness of our model, similar to previous work (Zang et al., 2021). For multi-modal dialog retrieval, we use ranking-based evaluation metrics such as recall n at k including *R@1*, *R@5* and *R@10* in accordance with prior studies (Zang et al., 2021; Shuster et al., 2018). These metrics measure whether the ground-truth textual or visual outputs

Task	Dataset	Metric	Previous SOTA	PaCE
Multi-Modal Intent Prediction	PhotoChat	F1-Score	58.9 (T5-3B)	**63.8** (+4.9)
	MMDialog	F1-score	75.5 (Divter)	**77.6** (+2.1)
Multi-Modal Dialog Retrieval	PhotoChat (T2I)	R@1	10.4 (SCAN)	**15.2** (+4.8)
	Image-Chat (I2T)	R@1	50.3 (TransResNet)	**51.9** (+1.6)
Multi-Modal Dialog State Tracking	MMConv	Acc.	18.0 (DS-DST)	**39.2** (+21.2)
	SIMMC2.0	Act-F1	96.3 (BART-large)	**97.1** (+0.8)
Multi-Modal Response Generation	MMConv	Comb.	32.2 (SimpleTOD)	**44.7** (+12.5)
	SIMMC2.0	BLEU	33.1 (BART-large)	**34.1** (+1.0)

Table 2: Experimental results on various multi-modal dialogue benchmarks. We compare PaCE with previous state-of-the-art models, including T5-3B (Raffel et al., 2020), Divter (Feng et al., 2022), SCAN (Lee et al., 2018), TransResNet (Shuster et al., 2018), BART-large (Lewis et al., 2019) and SimpleTOD (Hosseini-Asl et al., 2020).

are ranked among the top $k \in \{1, 5, 10\}$ positions among n candidate elements. For multimodal dialogue state tracking, we report *Categorical*, *Non-categorical* and *overall* scores as evaluation metrics following (Liao et al., 2021). To measure the quality of response generation, we employ BLEU (Papineni et al., 2002) as the evaluation metric for SIMMC2.0. For MMConv, we report a combined score (Comb.), which is computed via $(Inform + Success) \times 0.5 + BLEU$ as an overall evaluation measure as in (Mehri et al., 2019).

5.4 Quantitative Comparison

As shown in Figure 2 and Table 2, PaCE demonstrates state-of-the-art performances across a wide range of multi-modal dialogue tasks. Specifically, we have achieved a significant enhancement on the PhotoChat and MMConv dataset, with an improvement of 4.8 points in multi-modal dialog retrieval and 21.2 points in multi-modal dialog state tracking, respectively. It is worth noting that PaCE has a total parameter count of 338 million. In addition, since some experts may be idle during the execution of specific downstream tasks, the parameter size will further decrease for specific downstream tasks. Below, we provide a detailed analysis of the results for each sub-task dataset.

Multi-Modal Intent Prediction For the PhotoChat dataset, we report the performances of strong baselines as in (Zang et al., 2021), including ALBERT-base (Lan et al., 2019), BERT (Devlin et al., 2018), T5-base, and T5-3B (Raffel et al., 2020). For the MMDialog dataset, we adopt DE++, Divter (Feng et al., 2022), and ViLT (Kim et al., 2021) as our baseline models. As shown in Table 3, although some models such as T5-3B are much larger than ours, our model still achieves the best performance on all evaluation metrics.

	PhotoChat				MMDialog	
Model	**F1**	**Precision**	**Recall**		**Model**	**F1**
ALBERT-base	52.2	44.8	62.7		DE++	59.0
BERT-base	53.2	56.1	50.6		Divter	_75.5_
T5-base	58.1	_58.2_	57.9		-	-
T5-3B	_58.9_	54.1	_64.6_		-	-
ViLT	52.4	55.4	58.9		ViLT	55.8
PaCE	**63.8**	**63.3**	**68.0**		**PaCE**	**77.6**

Table 3: Multi-modal intent prediction results on PhotoChat and MMDialog.

Model	R@1	R@5	R@10	Sum(R@1,5,10)
BM25	6.6	15.4	23.0	45.0
DE*	9.0	26.4	35.7	71.1
VSE++	10.2	25.4	34.2	69.8
SCAN	10.4	27.0	37.1	74.5
VLMo	_13.8_	_30.0_	_39.4_	_83.2_
ViLT	11.5	25.6	33.8	71.0
PaCE	**15.2**	**36.7**	**49.6**	**101.5**

Table 4: Multi-modal dialogue retrieval on PhotoChat.

Multi-Modal Dialog Retrieval For PhotoChat, we compare PaCE with strong baselines reported in (Zang et al., 2021), including BM25 (Robertson et al., 2009), DE* (Zang et al., 2021), VSE++ (Faghri et al., 2017) and SCAN (Lee et al., 2018). We also adapted VLMo (Bao et al., 2021) and ViLT (Kim et al., 2021) to perform multi-modal dialog retrieval. The results on PhotoChat are reported in Table 4, PaCE achieves substantially better performance than the best performing baselines. For Image-Chat, we compare PaCE with TransResNet152 (Liao et al., 2021), VLMo and ViLT, and report baseline results as in Table 5. PaCE achieves the best results for image-to-text dialog retrieval with 3.0 improvement in terms of Sum.

Multi-Modal Dialog State Tracking For MMConv dataset, we compare PaCE with DS-DST(Zhang et al., 2019); for SIMMC2.0 dataset, we compare PaCE with GPT-2 (Radford et al., 2019), MTN (Le et al., 2019), BART-large and

Model	R@1	R@5	Sum(R@1,5)
TransResNet152	40.6	67.2	107.8
TransResNet152-IG-3.5B	50.3	75.4	125.7
VLMo	46.8	67.5	114.3
ViLT	48.4	70.0	118.4
PaCE	**51.9**	**76.8**	**128.7**

Table 5: Multi-modal dialog retrieval on Image-Chat.

Model	Categorical	Non-categorical	Overall
DS-DST	91.0	23.0	18.0
PaCE	**92.2**	**43.4**	**39.2**

Table 6: Multi-modal dialog state tracking performances on MMConv.

	Dialog State Tracking		Dialog Generation
Model	Slot F1	Act. F1	BLEU
GPT-2	81.7	94.5	19.2
MTN	76.7	93.4	21.7
BART-large	**88.3**	96.3	33.1
BART-base	82.0	95.2	29.4
PaCE	87.0	**97.1**	**34.1**

Table 7: Multi-modal dialog state tracking on SIMMC2.0. The evaluation metrics Slot F1 and Act. F1 are used to evaluate the dialog state tracking task, while BLEU is adopted for evaluating response generation.

Model	Inform	Success	BLEU	Comb.
SimpleTOD	14.6	9.2	20.3	32.2
PaCE	**34.5**	**13.9**	**22.0**	**44.7**

Table 8: Multi-modal response generation performances on MMConv.

BART-base (Lewis et al., 2019). The results on MMConv and SIMMC2.0 are reported in Table 6 and Table 7, respectively. PaCE can achieve the best results on most of the evaluation metrics. Notably, we observed that the PaCE achieves competitive results at smaller parameter scales than previous SOTA in SIMMC2.0 slot F1.

Multi-Modal Response Generation For the response generation task, we conduct experiments on SIMMC2.0 and MMConv datasets. For MMConv, we adopt the strong baseline SimpleTOD (Hosseini-Asl et al., 2020) implemented by (Liao et al., 2021). We summarize the experimental results of SIMMC2.0 and MMConv in Table 7 and Table 8, verifying the effectiveness of our model in both discriminative and generative tasks.

5.5 Ablation Study

Effectiveness of Pre-training Objectives To evaluate the effectiveness of each stage of pre-training, we conduct an ablation study by removing Stage I pre-training (PaCE$_{w/o\ \mathcal{L}^{I}_{stage}}$), removing

Stage II pre-training (PaCE$_{w/o\ \mathcal{L}^{II}_{stage}}$), removing Stage III pre-training (PaCE$_{w/o\ \mathcal{L}^{III}_{stage}}$), and removing both Stage II and Stage III (PaCE$_{only\ \mathcal{L}^{I}_{stage}}$). For a fair comparison, the experimental setup of the ablation study is consistent with that of the primary experiments, utilizing the same hyper-parameters and downstream fine-tuning strategy. The ablation test results on PhotoChat and Image-Chat are provided in Table 9. We can observe that image-text matching (Stage I) and image-context matching (Stage II) play the most important role in PaCE. This is within our expectation since Stage I and Stage II are the basis of the latter generation modeling (Stage III). It is no surprise that combining all three stages achieves the best performance on the experimental datasets. We also investigate the impact of $\mathcal{L}_{tca}$ by removing it from Stage II pre-training (denoted as PaCE$_{w/o\ \mathcal{L}_{tca}}$). We can observe that $\mathcal{L}_{tca}$ has a significant impact on the performance of PaCE in Stage II pre-training.

Effectiveness of Pre-training Data In addition, we also conduct an ablation study to verify the impact of different pre-training data on PhotoChat and Image-Chat datasets. We define the models that only use MultiNonDialog and MultiDialog for pre-training as PaCE$_{only\ MultiNonDialog}$ and PaCE$_{only\ MultiDialog}$, respectively. The ablation test results on PhotoChat and Image-Chat are provided in Table 10. We can observe that both MultiNonDialog and MultiDialog pre-training corpora contribute great performance improvement to PaCE. This is within our expectation since the MultiNonDialog data helps our model learn impressive image-text representations and their alignment, while the MultiDialog data encourages PaCE to capture the dialog context information.

Model	PhotoChat		Image-Chat	
	R@1	Sum(R@1,5,10)	R@1	Sum(R@1,5)
PaCE	**15.2**	**101.5**	**51.9**	**128.7**
PaCE$_{w/o\ \mathcal{L}^{I}_{stage}}$	10.7	74.3	46.5	117.8
PaCE$_{w/o\ \mathcal{L}^{II}_{stage}}$	12.0	74.8	48.5	119.5
PaCE$_{w/o\ \mathcal{L}^{III}_{stage}}$	15.0	100.8	51.2	127.3
PaCE$_{w/o\ \mathcal{L}_{tca}}$	13.2	95.9	49.7	125.6

Table 9: Ablation test results on the multi-modal dialog retrieval task by using different pre-training objectives.

Model	PhotoChat		Image-Chat	
	R@1	Sum(R@1,5,10)	R@1	Sum(R@1,5)
PaCE	**15.2**	**101.5**	**51.9**	**128.7**
PaCE$_{only\ MultiNonDialog}$	10.9	73.6	47.1	116.9
PaCE$_{only\ MultiDialog}$	10.7	74.3	46.2	117.3

Table 10: Ablation test results on the multi-modal dialog retrieval task by using different pre-training data.

6 Conclusion

In this paper, we proposed PaCE, a unified, structured, compositional multi-modal dialogue pre-training framework, which adopted a divide-and-conquer strategy. We first break down the complicated multi-modal dialogue generation task into several sub-capabilities, which could be learned in an easier way. Then, the solutions to the sub-capabilities were combined to obtain an effective and efficient solution to each downstream multi-modal dialogue task. Experimental results on eight benchmark datasets demonstrated that PaCE achieved new state-of-the-art performances.

Discussion

PaCE adopts a flexible model structure that decomposes complex multimodal dialogues into basic sub-capabilities. As a result, it can be trained progressively on different data and exhibits excellent expandability, making it applicable to new tasks. An additional advantage is that it aligns well with various attempts to enhance performance in terms of interpretability. However, we believe that there are still many aspects of PACE that are worth exploring. First is the exploration of incorporating additional modalities and investigating whether the self-attention layer can effectively handle a broader range of modalities for a unified representation. Another aspect worth exploring is the development of a more efficient approach for adapting multi-modal models to diverse downstream applications, eliminating the necessity to fine-tune all parameters of the model. Furthermore, given the substantial variations in the model networks employed for text generation and image generation in contemporary research, exploring the integration of multi-modal generation into a unified framework is a worthwhile endeavor.

Limitations

To better analyze the limitations of PaCE, we carry out an analysis of the errors made by PaCE on the PhotoChat and SIMMC2.0 test sets. We reveal several reasons for the errors, which can be divided into the following categories. **First**, since there are many similar images in the datasets, PaCE fail to distinguish some gold image from similar candidates. This may be because we do not design an explicit fine-grained reasoning module to capture the details of images and texts. For example, for the

context mentions "*I and my dad both have a camera*", our model can capture the entity "*camera*", but fails to reason the fact that there should be two cameras. One possible solution is to introduce a deep reasoning and comprehension strategy to empower the model with excellent reasoning ability. **Second**, due to the lack of fine-grained structural understanding of the images, the sentences generated by PaCE suffer from identifying the relative positions of entities. For example, PaCE may have difficulties recognizing the fact that the right side of a yellow shirt is black pants. This issue is particularly severe in SIMMC as there are many entities in the pictures and spatial descriptions of entities in the responses. One possible idea is to extract the relative positions of objects mentioned in the conversation as auxiliary data to guide the model's generation.

Acknowledgements

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. 2021. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*.

Feilong Chen, Xiuyi Chen, Can Xu, and Daxin Jiang. 2021. Learning to ground visual objects for visual dialog. *arXiv preprint arXiv:2109.06013*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and

Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

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*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.

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.

Binyuan Hui, Ruiying Geng, Qiyu Ren, Binhua Li, Yongbin Li, Jian Sun, Fei Huang, Luo Si, Pengfei Zhu, and Xiaodan Zhu. 2021. Dynamic hybrid relation exploration network for cross-domain context-dependent semantic parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13116–13124.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

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*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Jonáš Kulhánek, Vojtech Hudecek, Tomáš Nekvinda, and Ondrej Dušek. 2021. Augpt: Dialogue with pre-trained language models and data augmentation. *arXiv preprint arXiv:2102.05126*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv preprint arXiv:1907.01166*.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee, et al. 2022. Learning to embed multi-modal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 813–830.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216.

Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyun Myaeng. 2021. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. *arXiv preprint arXiv:2107.08685*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven CH Hoi. 2022a. Lavis: A library for language-vision intelligence. *arXiv preprint arXiv:2209.09019*.

13411

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. Mmconv: an environment for multimodal conversational search across multiple domains. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 675–684.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yuxing Long, Binyuan Hui, Fulong Ye, Yanyang Li, Zhuoxin Han, Caixia Yuan, Yongbin Li, and Xiaojie Wang. 2023. Spring: Situated conversation agent pretrained with multimodal questions from incremental layout graph. *arXiv preprint arXiv:2301.01949*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured fusion networks for dialog. *arXiv preprint arXiv:1907.10016*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10860–10869.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2020. Multi-modal open-domain dialogue. *arXiv preprint arXiv:2010.01082*.

Shuzheng Si, Shuang Zeng, and Baobao Chang. 2022. Mining clues from incomplete utterance: A query-enhanced network for incomplete utterance rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4839–4847.

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2021. Multi-modal dialogue response generation. *arXiv preprint arXiv:2110.08515*.

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.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

13412

Ze Yang, Wei Wu, Huang Hu, Can Xu, Wei Wang, and Zhoujun Li. 2021. Open domain dialogue generation with latent images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14239–14247.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *arXiv preprint arXiv:2108.01453*.

Jian-Guo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wan, Philip S Yu, Richard Socher, and Caiming Xiong. 2019. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. *arXiv preprint arXiv:1910.03544*.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2021. Mmchat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*.
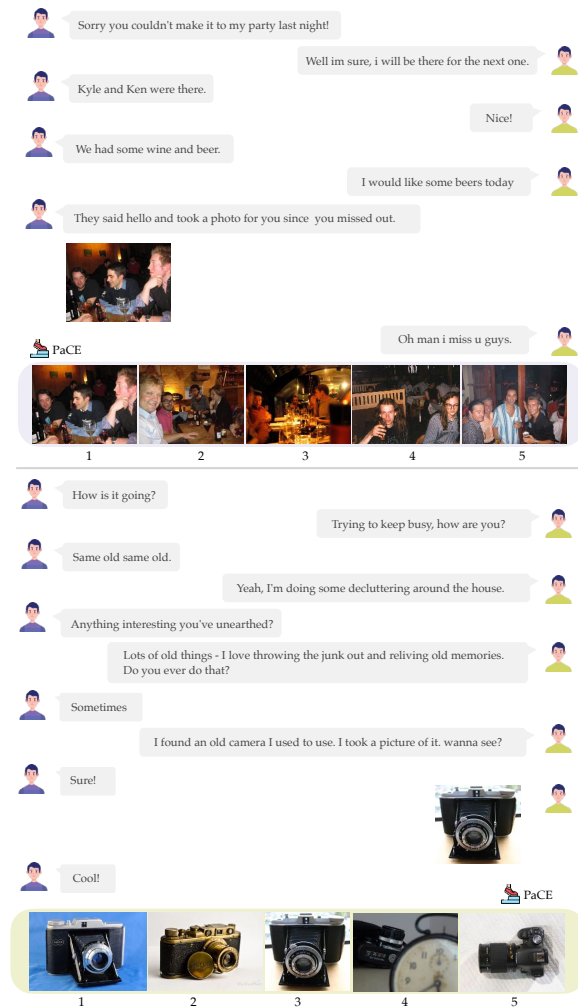
Figure 4: Two cases on the PhotoChat test set. For each dialog query, we show the top-5 ranked images from left to right.

A Case Study

To evaluate PaCE qualitatively, we choose two exemplary conversations from PhotoChat and ImageChat test sets, and illustrate the retrieved responses by PaCE in Figure 4 and Figure 5. Our PaCE model can retrieve highly relevant candidates to the conversation scenario. For the text-to-image (T2I) retrieval task, since the candidate images could be quite similar, it is challenging to retrieve the exact ground-truth image from the candidates. Although PaCE may not obtain the ground-truth image, we can still obtain the relevant candidate images.

Figure 5: Two cases on the Image-Chat test set. For each dialogue query, we show the top-5 ranked response from top to down.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ **A1.** Did you describe the limitations of your work?
We put the limitation after section 6.

☒ **A2.** Did you discuss any potential risks of your work?
There are no potential risks in our work.

☑ **A3.** Do the abstract and introduction summarize the paper's main claims?
We sumarize it in section 1.

☒ **A4.** Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

in Section 4.

☑ **B1.** Did you cite the creators of artifacts you used?
in Section 2.

☐ **B2.** Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☒ **B3.** Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ **B4.** Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ **B5.** Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ **B6.** Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In section 3.

C ☑ Did you run computational experiments?

In section 5.

☑ **C1.** Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In section 5.2

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In section 5.2

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In section 5.3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.