Knowledge Transfer with Visual Prompt in Multi-modal Dialogue Understanding and Generation

Minjun Zhu[1,2]*, Yixuan Weng[1]*, Bin Li[3], Shizhu He[1,2], Kang Liu[1,2], Jun Zhao[1,2]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] College of Electrical and Information Engineering, Hunan University
zhuminjun2020@ia.ac.cn, wengsyx@gmail.com, libincn@hnu.edu.cn,
{shizhu.he, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

Visual Dialogue (VD) task has recently received increasing attention in AI research. VD aims to generate multi-round, interactive responses based on the dialog history and image content. Existing textual dialogue models cannot fully understand visual information, resulting in a lack of scene features when communicating with humans continuously. Therefore, how to efficiently fuse multi-modal data features remains to be a challenge. In this work, we propose a knowledge transfer method with visual prompt (**VPTG**) fusing multi-modal data, which is a flexible module that can utilize the text-only seq2seq model to handle VD tasks. The VPTG conducts text-image co-learning and multi-modal information fusion with visual prompts and visual knowledge distillation. Specifically, we construct visual prompts from visual representations and then induce sequence-to-sequence (seq2seq) models to fuse visual information and textual contexts by visual-text patterns. Moreover, we also realize visual knowledge transfer through distillation between two different models' text representations, so that the seq2seq model can actively learn visual semantic representations. Extensive experiments on the multi-modal dialogue understanding and generation (MDUG) datasets show the proposed VPTG outperforms other single-modal methods, which demonstrate the effectiveness of visual prompt and visual knowledge transfer.

1 Introduction

Cross-modal understanding between vision and language has become a challenging field in natural language processing and computer vision. With the rapid development of deep neural networks, researchers have made rapid progress in a series of visual language tasks, including moment localization with natural language (Zhang et al., 2019a, 2020; Tan et al., 2021; Li et al., 2022b), image captioning (Vinyals et al., 2015; Chen et al., 2017; Anderson et al., 2017), visual question answering (Tang et al., 2018; Chen et al., 2020; Sheng et al., 2021), etc. The visual dialogue task (Das et al., 2017) aims to perform multiple rounds of interactive dialogue based on dialogue history and image content.

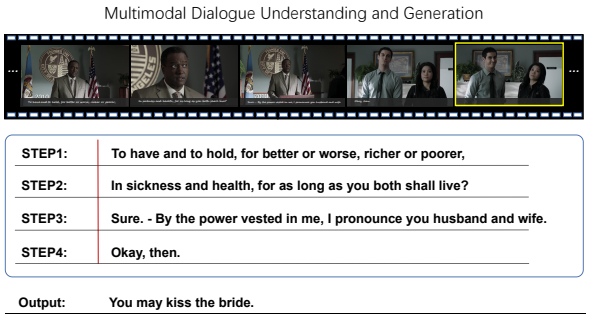

Figure 1: Description of the Multi-modal Dialogue Understanding and Generation (MDUG) task. From step1 to step 3, the video is about a priest, and the subtitles are snippets of wedding vows. For the response generation of step 4, supposing that only dialogue text context was taken, the previous dialog text: "OK, then" is inadequate for generating the expected output: "you may kiss the bride."

Dialogues with multi-modal contexts (visual and textual) are becoming more and more general in daily life (Baltrušaitis et al., 2018), such as communicating messenger tools (e.g. Facebook, WeChat). Compared with visual question answering, Visual Dialogue (VD) tasks not only require answering questions according to visual information but also require a deep understanding of multiple rounds of historical dialogues (Schwartz et al., 2019b; Gan et al., 2019; Chen et al., 2022). In the visual dialogue task, researchers have put forward a lot of relevant datasets, the *GuessWhat?!* (de Vries et al., 2016) and the *Visdial* (Das et al., 2017) set up visual dialog data sets for images. The MDUG (Wang et al., 2022b) is based on video scenes to generate coherent textual responses.

*These authors contributed to this paper equally.

Proceedings of the 29th International Conference on Computational Linguistic, pages 8–19
October 12–17, 2022.

In this work, we mainly focus on video visual dialogue such as the Multi-modal Dialogue Understanding and Generation (MDUG) dataset (Wang et al., 2022b). Compared to image captioning and image visual dialogue, it requires modeling long-distance image sequences, which is more challenging and practical. The MDUG task proposes a multi-modal dialogue task in the video field. It needs the system to generate a response of the current frame based on multi-modal video scene and historical dialogue information, where historical video clips frame and text captions are mapped one-to-one. The video clips and visual images have much abundant and useful information about the plot development. It is easy to pick up on their movements and expressions from visual information. For example, in the last frame of Figure 1. On the one hand, from the body movements of people such as they gradually face each other and a smile on the man's face, we can observe that the man is going to kiss his bride, so models can infer the "kiss" action in generated response. On the other hand, from the wedding vows context, it's easy to infer their roles as bride and groom. Therefore, this example demonstrates the importance of combining images and texts for the MDUG task.

Although much attention has been drawn to dialogue tasks (Das et al., 2017), neural models have shown impressive performance gains in textual dialogue tasks. But existing text-only dialogue methods still have limitations in handling video dialogue tasks in multi-modal scenarios, which may hinder further advancement in this direction. In text-only dialogue tasks, more and more text generation models are pre-trained in the large-scale corpora with the development of pre-trained language models (Brown et al., 2020; Shao et al., 2021). Most of the dialogue pre-training models are based on transformers through pre-training in large-scale dialogue texts and using a large number of encoder and decoder layers (Gu et al., 2022; Zhou et al., 2021; Bao et al., 2021). This can improve the consistency between the generated context and context and the fluency of the generated text. But the bigger challenge is based on the non-homogeneity of the input text-image multi-modal information and the output text information besides challenges in the text-only task in multi-modal dialogue generation tasks.

How to understand and integrate the multi-modal information, and comprehensively perform text generation remains to be an unsolved and important problem. Many efforts have been made to realize a reliable and accurate multi-modal dialogue understanding and generation in similar tasks such as image captioning and video question answering (Fukui et al., 2016; Sharma et al., 2020; Das et al., 2017; Shrestha et al., 2019). However, the methods adopted in that work cannot be directly generalized to the video visual dialogue task, and the video visual dialogue task requires multi-level modeling in a large number of sequence images and dialog history at the same time (Schwartz et al., 2019a).

To take a significant step in this direction and fully utilize seq2seq models' capability, we propose a Visual Prompt Text Generate (**VPTG**) method that can directly provide visual assistance training for multi-modal language models to tackle the above challenges. The VPTG framework can efficiently generate dialogue response that is coherent to both visual images and text dialogue. To model text-image mapping in the same representation space, we adopt CLIP contrastive training to conduct co-learning of image-caption pairs through a pre-trained language model (Liu et al., 2021a). We also use the visual prompt to fuse image visual information into text features. In the training stage, we input the "image" and "answer text" into the CLIP (Radford et al., 2021), and input the "image" feature vector as a visual prompt into the seq2seq model. In addition, to improve the visual modeling ability of language models, we conduct visual knowledge transfer by transferring visual representations to visual prompt and using it to prompt the seq2seq model modeling multi-modal data. Specifically, the "answer text" feature is also provided to the encoder output "[CLS]" vector of the seq2seq model for distillation. We also ask the sequence-to-sequence (seq2seq) model to actively learn visual semantic representations. For efficient training, we adopt an end-to-end training architecture.

In the prediction stage, we only use the image as the input of the CLIP and get the visual prompt, and then perform multi-level learning from visual information to textual information. In the VPTG, we perform efficient representation, co-learning, and fusion of multi-modal information. Extensive experimental results show that the VPTG method consistently outperforms all baseline schemes in the MDUG task, showing the effective ability of the method to make better use of textual and visual information to generate high-quality multi-modal

dialogue responses.

In summary, our contributions are as follows:

- In this work, we focus on the video visual dialogue task. To the best of our knowledge, this is the very first attempt to apply the visual prompt for solving the video dialogue response generation task.
- We present a useful method, which can be used in almost all seq2seq models. And it conducts visual prompts and visual knowledge transfer to jointly learn images and text, and effectively generate a response. We explore the task with multi-modal information representation, co-learning, and fusion.
- Extensive experiments are performed to examine the effectiveness of the proposed VPTG on the MDUG dataset, in which we achieve state-of-the-art performances.

2 Related work

2.1 Visual Dialogue Task

With the progress of human-robot interaction technology, more and more dialogue tasks emphasize user-friendliness and ethical safety (Zhang and Zhao, 2021). A dialogue system mainly includes two parts: (1) understanding the history of dialogue; (2) Response in natural language.

The Visual Dialogue (VD) task require agents to have meaningful dialogue with humans in multi-modal scenes (Das et al., 2017; Dalu et al., 2019; Li et al., 2021; Wang et al., 2022b). It is more complex than traditional visual tasks (such as Object Detection (Ren et al., 2015), Image Retrieval (Kalantidis et al., 2015)). In the VD task, given some frame or a video clip, a dialog history context, the agent has to ground in image and text, infer context from history, and generate text response accurately. It requires multi-dimensional modeling based on visual information to generate accurate descriptions, which has been used to help visually impaired people better understand the visual content of the environment. The **MDUG** dataset is a VD dataset that aims to generate an interactive response based on the image captions context history and video clips image content. The traditional multi-modal fusion method first uses the visual model to extract the image features and then uses the neural network such as LSTM (Hochreiter and Schmidhuber, 1997) to fuse the information between different modes. In recent years, many methods have been committed to more comprehensive information fusion (Vinyals et al., 2014), such as MHCIAE (Lu et al., 2017) used discriminative learning to migrate knowledge into dialogue generation. ReDAN (Gan et al., 2019) conducted visual dialogue through multi-step reasoning. UTC (Chen et al., 2022) unified the discriminative and generation of Visual Dialogue tasks based on the framework of contrastive learning. Different from previous works, the **VPTG** adopts a more flexible and widely applicable framework that can be integrated with various single-modal pre-trained language models to learn vision-language interactions by taking visual prompt and visual knowledge transfer, which deeply captures the relations between image and texts to mutually reinforce dialogue response generation.

2.2 Pre-Trained Language Model

There are also pre-trained models promising in the visual-language field (Murahari et al., 2019; Wang et al., 2020; Ye et al., 2022). Most of the popular approaches employ an encoder-decoder architecture for visual dialog. The encoder aims at encoding the image and text to fused features, and two separate decoders are employed for ranking and generating respectively. Among them, a variety of attention mechanism-based approaches are proposed to learn the interactions between the image, the answers, and the dialog history in the discriminative setting. The 3D ConvNet was pre-trained on the Kinetics dataset (Carreira and Zisserman, 2017). The CLIP (Radford et al., 2021) and Wenlan (Huo et al., 2021) models are image-text pair pre-trained models, which are pre-trained by learning to map text and image to the same vector space. The OFA (Wang et al., 2022a) is a unified model adopting multi-modality pre-training with multi-tasking training objectives. It transforms all multi-modal tasks into sequence-to-sequence (seq2seq) tasks, which realizes the state-of-the-art performance in multiple visual-language tasks.

2.3 Prompt Tuning

How to make better use of pre-trained models has become a concerning problem (Han et al., 2021b). Prompt tuning is a new NLP paradigm used to solve the downstream tasks of the pre-trained model.In the field of multi-modality, increasing methods adopt prompt tuning to learn the aligned features between different modalities. CPT (Yao et al.,

2022) uses color (visual feature) as a bridge to recover masked tokens from cross-modal content, narrowing the gap between pre-training and downstream tasks. The VPTSL (Li et al., 2022a) formulates the natural language video localization task as an extraction reading comprehension task by introducing the discrete visual prompt. And, it implements a new state-of-the-art on the MedVidQA (Gupta et al., 2022) datasets.

The VPTG solves the defect of incomplete utilization of visual features. It also performs visual prediction tasks by $\mathcal{L}_{\mathrm{KL}}$ compared with these prompt methods. This can make the model more fully understand the visual semantics, so as to better multi-modal modeling.

3 Datasets

The multi-modal Dialogue Understanding and Generation task (Wang et al., 2022b) is required to generate a dialogue agent for the next sentence based on the multi-modal scene and the previous dialogue process. This task needs to model the semantics of the session and the scenario of the session. The task provides the multi-modal video of dialogue content and scene. Its ultimate goal is to generate agent replies that meet the context and are related to the video scene.

The videos and dialogues for this task are crawled from online TV series. The dataset is split into a training set, a validation set, and a test set. Each example includes a dialogue session as well as the associated video clip, which is a sequence of frames. The frames from the videos have been downsampled to 3fps.

It is composed of 43,895 videos with 1,100,242 utterances. Each video has an average of 25.07 utterances. We follow the official data split, where 1,000,079, 50,032, and 50,131 utterances are used for training, validation, and testing, respectively.

4 The Proposed Method

We propose the visual prompt Text Generate (VPTG) framework for the multi-modal Dialogue Understanding and Generation (MDUG) task, whose ultimate goal is to generate a response that is coherent to the dialogue context and relevant to the video context. The Figure 2 illustrates the architecture of VPTG. It is challenging to directly generate the dialogue response according to multi-modal data. To tackle this challenge of data alignment and fusion between image and text, we split the MDUG task into two simultaneous modules: (1) the visual predictor module is first used to generate **visual prompt** (Section 4.1) by jointly training an image encoder and a text encoder and fusion image information into a text representation. (2) The text predictor conducts **Visual Knowledge Transfer** (Section 4.2) to guarantee response generation with information alignment between text and image.

4.1 Visual Prompt

The visual prompt method was proposed in the Visual Predictor module. In this module, we aim at learning multi-modal feature representation and constructing visual prompts to reinforce semantic modeling.

In the MDUG task, an example includes a dialogue session and the associated video clip which is a sequence of frames (3 frames per second). In the VPTG, we input the last frame of video I and a corresponding next textual response T corresponding at a time. Because image and text are heterogeneous data, we leverage the CLIP (Radford et al., 2021) to model joint representations of image and text. For multi-modal data, joint representations are projected to the same space using all of the modalities as input. The CLIP (Radford et al., 2021) is a visual-language pre-training model that learns both visual and language representations by predicting the correct pairings of a batch of $\{\text{image}, \text{text}\}$ training examples. In our model, for the current frame image and the next textual response, we utilize an image encoder to get visual prompt $\mathcal{V}_{\text{image}} \in \mathbb{R}^k$ and a text encoder to get $V_{\text{text}} \in \mathbb{R}^k$, they are jointly trained to respectively map the input image and text into a unified representation space. We adopt contrastive learning as its training objective. We use L_{CL} to close the semantic distance of image-text pairs, where ground truth image-text pairs are regarded as positive samples $\mathbf{X}^+ = \{\mathbf{x}_i^+\}_{i=1}^n$, and mismatched image-text pairs constructed as negative ones $\mathbf{X}^- = \{\mathbf{x}_i^-\}_{i=1}^m$.

$$L_{CL} = -\sum_{i=0}^{n}\left[\log \frac{\mathrm{Sim}\left(x_i, x_i^+\right)}{\mathrm{Sim}\left(x_i, x_i^+\right) + \sum_{j=1}^{m}\mathrm{Sim}\left(x_i, x_j^-\right)}\right]$$
$$\mathrm{Sim}(x_i, x_j) = \exp\left(f(x_i)^T f\left(x_j\right)\right) \quad (1)$$

4.1.1 Prompt Designing

The information coming from text and image modalities may have varying predictive power and noise topology (Baltrušaitis et al., 2018). After learning joint representations of image-text pairs,

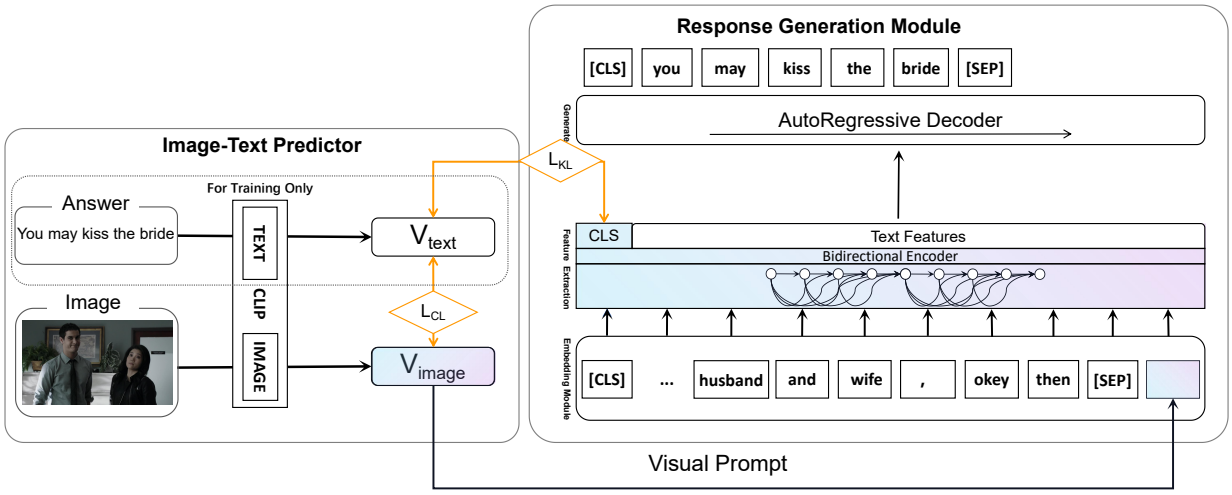

Figure 2: The architecture of the proposed method. In the training stage, we input the "image" and "answer text" into two separate encoders of CLIP, and input the image feature vector as a visual prompt into the seq2seq model. In addition, the answer eigenvector is also provided to the encoder output "[CLS]" vector of the seq2seq model for distillation. In the prediction stage, we only use the image as the input of the CLIP and get the visual prompt.

we conduct visual prompt learning. Unlike traditional visual prompt Tuning methods aiming to finetune large-scale Transformer modules with a small amount of task-specific learnable parameters, we construct the visual prompt to fuse visual modality into text modeling and generation, which can also be trained end-to-end.

We adopt the visual image representation as the visual token for prompting the pre-trained language model. Specifically, the image representation V_{image} was transferred to the same dimension as the input text tokens as a visual prompt.

$$\mathcal{P}_{\text{t}} = Linear(\mathcal{V}_{\text{image}}) \tag{2}$$

where $\mathbf{P}_{\text{t}} \in \mathbb{R}^d$,d is the dimension of text predictor encoder embedding; $Linear$ is a single feed-forward layer.

4.1.2 Prompt Tuning

Intuitively, the visual prompt $\mathcal{P}_{\text{t}}$ is used as the visual token which concatenates with the text dialogue sentence and the last video frame image. The "[CLS]" is positioned at the head of the input token, while the prompt $\mathcal{P}_{\text{t}}$ is used as the trigger to model and generate a response. After concatenation, the embedding module is adopted for learning the features in the same vector space. On the one hand, the visual prompt covers the non-verbal part that the text token lacks. On the other head, the visual prompt is supervised by the visual frames, where some visual features can be the extra knowledge for the pre-trained model when fine-tuning.

$$\mathbf{P} = \mathbf{Embedded}\left([\text{CLS}]\mathbf{Text}[\text{SEP}]\right)\mathbf{Concat}\mathcal{P}_{\text{t}} \tag{3}$$

4.2 Visual Knowledge Transfer

The text predictor module is based on the seq2seq Transformer model (Vaswani et al., 2017a). The Transformer is Encoder-Decoder architecture, which is proved to be outstanding for text generation. The encoder produces a global contextual representation based on multi-modal representation fusion, and the decoder will use the multi-head attention mechanism to fuse encoder information, and then generate the final frame predicted response token by token. To make information alignment, we propose **Visual knowledge transfer** to distil knowledge by cross-attention. This thought has been proved to perform better multi-modal information fusion in the textual question answering field (Izacard and Grave, 2020).

4.2.1 Text Encoder Distill Learning

In text predictor, each $\mathbf{P}$ constructed in Visual Prompt is given as input to a seq2seq model encoder.

$$V_P = \text{Encoder}_{seq2seq}(\mathbf{P}) \tag{4}$$

Let $V_{CLS}^{seq2seq} \in \mathbb{R}^d$ be the [CLS] token's representation of the encoded query V_P, it models the whole representation containing dialogue text and visual prompt in the bidirectional encoder. We will

assume that the last hidden state output among two encoders and text can be defined as $p_1(t \mid p)$ and $p_2(t \mid z)$. There are two transformer encoders in the VPTG, where we call the visual predictor encoder as $\mathbf{Encoder}_1$, the text predictor encoder as $\mathbf{Encoder}_2$.

$$p_1(t \mid p) \propto V_{text}^{CLIP}, \quad p_2(t \mid z) \propto V_{CLS}^{seq2seq} \quad (5)$$

where t is input dialogue text, p is the input frame image; z is the visual prompt according to p; $V_{text}^{CLIP} \in \mathbb{R}^k$ is the representation of image in the visual predictor. The p_1 represent the $\mathbf{Encoder}_1$, and the p_2 represent the $\mathbf{Encoder}_2$. We close the gap between $V_{CLS}^{seq2seq}$ and V_{text}^{CLIP} by minimizing the KL-divergence. This aims at training the response generator ($\mathbf{Encoder}_2$) with visual knowledge information from the image-text predictor ($\mathbf{Encoder}_1$).

$$\begin{aligned}
\mathcal{L}_{\mathrm{KL}}^{0}(\theta, \mathcal{P}_{\mathrm{t}}) &= D_{KL}(V_{CLS}^{seq2seq}(x) || w_0 V_{text}^{CLIP}(x)) \\
\mathcal{L}_{\mathrm{KL}}^{1}(\theta, \mathcal{P}_{\mathrm{t}}) &= D_{KL}(w_0 V_{text}^{CLIP}(x) || V_{CLS}^{seq2seq}(x)) \\
\mathcal{L}_{\mathrm{KL}}(\theta, \mathcal{P}_{\mathrm{t}}) &= \frac{1}{2} \sum_{x \in \mathcal{X}} \left(\mathcal{L}_{\mathrm{KL}}^{0}(\theta, \mathcal{P}_{\mathrm{t}}) + \mathcal{L}_{\mathrm{KL}}^{1}(\theta, \mathcal{P}_{\mathrm{t}}) \right) \quad (6)
\end{aligned}$$

where $\mathcal{X}$ is the training set of all image-text pairs. $w_0 \in \mathbb{R}^{d \times k}$ is a trainable weights vector. The text predictor encoder ($\mathbf{Encoder}_2$) is trained simultaneously by the response generation task. We take the formula above to perform visual knowledge distill learning. In training $\mathcal{L}_{\mathrm{KL}}$, it performs gradient decoupling (stop-gradient operator) for $V_{text}^{CLIP}(x)$ and $\mathbf{Encoder}_1$. This visual knowledge distill learning method requires the seq2seq model (or $\mathbf{Encoder}_2$) to actively learn visual semantic representation, so as to increase the model's perception of visual signals and avoid ignoring information of visual prompt.

4.2.2 Response Generation

Finally, we generate responses with the seq2seq model's decoder. We define L_{gen} as the autoregressive loss.

$$L_{gen} = -\sum_{n=1}^{N} p(y_i) \log \frac{\exp(y_i)}{\sum_{n=1}^{N} \exp(y_i)} \quad (7)$$

where y_i is the i-th generated token by the language model. N is the size of the target vocabulary.

4.3 Training and Inference

Combining the above derivations, our training objective that we seek to minimize for response becomes:

$$\mathcal{L} = \mathcal{L}_{\mathrm{KL}} + \lambda \mathcal{L}_{gen} + \gamma \mathcal{L}_{CL}, \gamma \in \mathbb{R}, \lambda \in \mathbb{R}. \quad (8)$$

We jointly train the visual predictor and text predictor as an end-to-end training approach.

For inference, we first encode the input image-text pairs by the visual predictor, then construct the visual prompt to fuse multi-modal representation. The text predictor can generate predicted responses after concatenation between the text tokens and the visual token.

5 Experiments

In this section, we will introduce the evaluation indicators and experimental settings. Then we compare VPTG with the existing dialogue generation technology and ablation experiments to prove the effectiveness of our method.

5.1 Evaluation Metrics

Following prior work (Chen et al., 2015; Laokulrat et al., 2016; Pasunuru and Bansal, 2017; Liu et al., 2021b), we use a variety of evaluation indicators, which can evaluate the generation quality of sentence level and word level at the same time, and show the detailed performance of the system more comprehensively. We adopt "BLEU" (Papineni et al., 2002), "ROUGE" (Lin, 2004), "METEOR" (Denkowski and Lavie, 2014) and "CIDER" (Vedantam et al., 2015) as the evaluation metrics, which can assess the quality of visual dialogue generation, including fidelity and diversity.

5.2 Implementation Details

In order to compare the functions of the system more fairly, we follow the setting of the baseline scheme and only compare whether to add the VPTG module. In recent years, natural language processing significant progress has been achieved (Han et al., 2021a; Qiu et al., 2020) due to the introduction of Pre-trained Language Model (Peters et al., 2018; Devlin et al., 2019; Radford and Narasimhan, 2018). Therefore, more and more methods begin to introduce the pre-trained language model in the dialogue generation task (Zhang et al., 2019b; Adiwardana et al., 2020; Roller et al., 2021b; Thoppilan et al., 2022; Gu et al., 2022).

For all methods, we use the same CLIP [1] (Radford et al., 2021) model as feature extraction It

[1] https://huggingface.co/openai/clip-vit-base-patch32

Models		BLEU-1	ROUGE-L	METEOR	CIDEr	Avg
Random Mode		4.81	3.92	2.21	2.42	3.34
BART-base(Lewis et al., 2019) (2019)	Originally	5.02	4.35	2.54	3.75	3.92
	Fintune	5.74	6.10	3.87	4.11	4.96
	With VPTG	**6.12**	**6.52**	**4.01**	**4.35**	**5.25(0.29↑)**
T5-base (2020)	Originally	2.78	4.21	2.33	**1.20**	2.63
	Fintune	2.94	4.44	2.81	0.58	2.69
	With VPTG	**3.24**	**5.12**	**2.98**	0.89	**3.06(0.37↑)**
Blender-400M (Roller et al., 2021a)(2021)	Originally	6.03	7.69	5.43	3.51	5.67
	Fintune	7.01	8.73	6.05	5.85	6.91
	With VPTG	**7.55**	**9.15**	**6.49**	**6.61**	**7.45(0.54↑)**

Table 1: Performance comparison of the variants methods on MDUG dataset. We highlight the best score in each column in **bold**, and the second best score with underline. We also show the improvement between first place and second place.

Case Study	BLEU-1	ROUGE-L	METEOR	CIDEr	Avg
Baseline	7.01	8.73	6.05	5.85	6.91
W/O $\mathcal{L}_{\mathrm{KL}}$	7.25	8.91	6.24	7.12	7.38
W/O Visual-Feature	7.10	8.79	6.34	6.01	7.06
W/O visual prompt	6.45	8.10	5.78	5.62	6.49
VPTG	**7.55**	**9.15**	**6.49**	**6.61**	**7.45**

Table 2: We conduct the ablation study to analyze the performance of the VPTG on the Blender-400M model, where we use the same parameters to train the model and report the highest score.

has 8 attention heads and 12 layers, and its hidden size is 512. For the seq2seq model, we all use the base size model for testing. And for the remaining settings, we follow the original code.

We train the model using the Pytorch[2] (Paszke et al., 2019) on the NVIDIA RTX3090 GPU and use the hugging-face[3] (Wolf et al., 2020) framework. We use the AdamW (Loshchilov and Hutter, 2018) as the optimizer and the learning rate is set to 1e-5 with the warm-up (He et al., 2016). The batch size is 24. We set the maximum length of 512 (we set the max length as 128 for Blender, because it supports up to 128 lengths of input), and deleted the excess. We use the linear decay of the learning rate and gradient clipping of 1e-6. The dropout (Srivastava et al., 2014) of 0.1 is applied to prevent overfitting. The detailed experimental settings are shown in **Table 1**.

All hyperparameters are optimized on the Valid set. In all our experiments, at the end of each training phase, we will test the effective data set and select the highest model (mainly depending on BLEU) in the test data set for prediction. We report the results in the test data set. We repeated the experiment three times and reported the average score.

[2] https://pytorch.org

[3] https://github.com/huggingface/transformers

5.3 Comparison with State-of-the-Art Methods

In the MDUG dataset, we compared the baseline scheme with the existing dialogue generation.

BART (Lewis et al., 2019) uses a standard seq2seq transformer (Vaswani et al., 2017b) structure. Its structure is very simple, which can be seen as a combination of BERT (Devlin et al., 2018) and GPT (Radford and Narasimhan, 2018). In the pre-training stage, it destroys the original text by randomly disrupting the order of the original sentences and adding mask tags. After that, the BART (Lewis et al., 2019) reconstructs the original text by denoising it. The BART (Lewis et al., 2019) achieves the best performance in translation and summary tasks that need to be generated.

T5 treats all tasks as text-to-text tasks. It is different from the BART (Lewis et al., 2019) in that the pre-training stage only requires the decoder to recover the masked part without full-text recovery. It has even surpassed the human level in many natural language tasks (Wang et al., 2018, 2019).

Blender (Roller et al., 2021a) is a pre-training model in the chat field. It carries out pre-training in a large number of dialogues, which improves the dialogue fluency of the model. It can provide users with interesting chat preferences, display personality, and so on. Blender can maintain consistent personality attributes in the dialogue and surpasses the existing models in terms of participation and

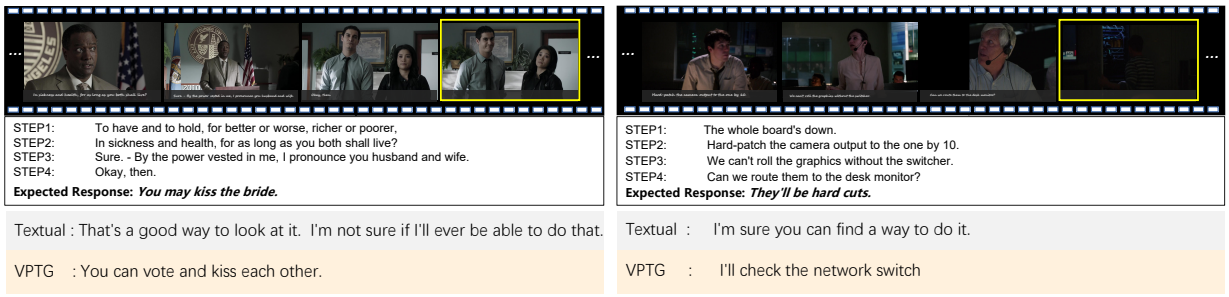

Figure 3: Examples of the generated results.

humanization indicators.

5.4 Experimental Result

We report the performance of the model in **Table 2**. The "Originally" refers to the use of the original pre-training model for a zero-shot generation. The "Finetune" means that we fine-tune the data set and select the highest score to test in the test. The "With VPTG" means that we have modified the structure of the model and added the VPTG module based on the existing language model, which enables us to give the visual ability to the language model that has never seen an image.

It is not difficult to find that, other models have poor zero-shot effects in the field of dialogue except the Blender. This is because the T5 model and the Bart model are pre-trained in a large-scale general corpus, which is difficult to migrate directly to the field of dialogue. Even if these models are fine-tuned, the effect is still insufficient, even worse than the result of random selection. This shows that Visual Dialogue tasks have strong open attributes and need to use more features.

After the VPTG is added to the model, the CLIP can provide visual semantic features. This makes the seq2seq model have a more comprehensive perceptual performance. It can analyze the overall scene and generate dialogue text more in line with the scene. In the "With VPTG" of Table 1, the performance of all models has been significantly improved. This shows the effectiveness of the VPTG module.

5.5 Ablation Study

In **Table 3**, we can see some performance comparisons. We further carry out care learning in Blender (Roller et al., 2021a), which is the best pre-trained model in MDUG tasks (Wang et al., 2022b). It can fully show the effect differences brought by different methods.

First, we try to cancel the $\mathcal{L}_{\mathrm{KL}}$ loss, which means that we no longer require the model to predict the actual video scene. This may lead to the lack of understanding of the scene in the model so that the generated text lacks the modelling of the scene.

After cancelling the visual feature, we will no longer provide the video feature vector of the current scene. This may make the model lack visual semantic features and cause the omission of environmental scenes.

We tested the use of dot products to integrate visual features into the embedding matrix of the seq2seq model, but the effects decreased significantly. We believe that if we do not use the visual prompt to provide visual features, the direct dot product will cause the catastrophic forgetting problem of the pre-training language model. It will destroy the original semantic understanding ability of the pre-training language model and become a kind of noise interference through the fusion of direct dot product feature vectors.

5.6 Case Study

In Figure 3, we select two examples to show. We can see that the VPTG model can better model scene information and generate text with specific visual semantics than the single modal language pre-training model. Compared with the single model, the VPTG has higher fluency in the field of dialogue. This fully shows that the VPTG can deeply mine visual signals.

6 Conclusions

In this paper, we proposed a new visual knowledge fusing paradigm that provides the pre-trained language generation model with the visual prompt. The VPTG module is flexible and can support almost all seq2seq models to be used in multi-modal dialogue generation tasks. It realizes the language model's understanding of visual infor-

mation by transforming visual features into embedding prompts. We have conducted vast experiments on the task of multi-modal Dialogue Understanding and Generation. The VPTG outperforms all other baselines in MDUG tasks for these experiments, which reflects the effectiveness of the proposed method.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *arXiv: Computation and Language*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and visual question answering. *computer vision and pattern recognition*.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xu Xinchao, Yingzhan Lin, and Zheng-Yu Niu. 2021. Plato-xl: Exploring the large-scale pre-training of dialogue generation. *arXiv: Computation and Language*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition*.

Cheng Chen, Yudong Zhu, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, and Xiaodong Gu. 2022. Utc: A unified transformer with inter-task contrastive learning for visual dialog.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. *computer vision and pattern recognition*.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *computer vision and pattern recognition*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Guo Dalu, Xu Chang, and Tao Dacheng. 2019. Image-question-answer synergistic network for visual dialog. *computer vision and pattern recognition*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. *computer vision and pattern recognition*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2016. Guesswhat?! visual object discovery through multi-modal dialogue. *computer vision and pattern recognition*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. 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. *north american chapter of the association for computational linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *empirical methods in natural language processing*.

Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. *meeting of the association for computational linguistics*.

Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Jie Tang, and Minlie Huang. 2022. Eva2.0: Investigating open-domain chinese dialogue systems with large-scale pre-training.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2022. A Dataset for Medical Instructional Video Classification and Question Answering. *arXiv preprint arXiv:2201.12888*.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021a. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021b. Pre-trained models: Past, present and future. *AI Open*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*.

Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, Zongzheng Xi, Yueqian Yang, Anwen Hu, Jinming Zhao, Ruichen Li, Yida Zhao, Liang Zhang, Yuqing Song, Xin Hong, Wanqing Cui, Dan Yang Hou, Yingyan Li, Junyi Li, Peiyu Liu, Zheng Gong, Chuhao Jin, Yuchong Sun, Shizhe Chen, Zhiwu Lu, Zhicheng Dou, Qin Jin, Yanyan Lan, Wayne Xin Zhao, Ruihua Song, and Ji-Rong Wen. 2021. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv: Computer Vision and Pattern Recognition*.

Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*.

Yannis Kalantidis, Clayton Mellina, and Simon Osindero. 2015. Cross-dimensional weighting for aggregated deep convolutional features. *european conference on computer vision*.

Natsuda Laokulrat, Sang Phan, Noriki Nishida, Raphael Shu, Yo Ehara, Naoaki Okazaki, Yusuke Miyao, and Hideki Nakayama. 2016. Generating video description using sequence-to-sequence model with temporal attention. *international conference on computational linguistics*.

Michael Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *meeting of the association for computational linguistics*.

Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. 2022a. Towards visual-prompt temporal answering grounding in medical instructional video. *arXiv preprint arXiv:2203.06667*.

Bin Li, Yixuan Weng, Fei Xia, Bin Sun, and Shutao Li. 2022b. Vpai_lab at medvidqa 2022: A two-stage cross-modal fusion method for medical instructional video classification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 212–219.

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. *IEEE Transactions on Audio, Speech, and Language Processing*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv: Computation and Language*.

Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. 2021b. Cptr: Full transformer network for image captioning. *arXiv: Computer Vision and Pattern Recognition*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. *neural information processing systems*.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *european conference on computer vision*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2017. Reinforced video captioning with entailment rewards. *empirical methods in natural language processing*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward

Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *CoRR*, abs/2003.08271.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021a. Recipes for building an open-domain chatbot. *conference of the european chapter of the association for computational linguistics*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021b. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. 2019a. A simple baseline for audio-visual scene-aware dialog. *computer vision and pattern recognition*.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G. Schwing. 2019b. Factor graph attention. *computer vision and pattern recognition*.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv: Computation and Language*.

Himanshu Sharma, Manmohan Agrahari, Sujeet Kumar Singh, Mohd Firoj, and Ravi Kumar Mishra. 2020. Image captioning: A comprehensive survey. In *2020 International Conference on Power Electronics IoT Applications in Renewable Energy and its Control (PARC)*, pages 325–328.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. *arXiv: Computer Vision and Pattern Recognition*.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2019. Answer them all! toward universal visual question answering models. *computer vision and pattern recognition*.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Yi Tan, Yanbin Hao, Xiangnan He, Yinwei Wei, and Xun Yang. 2021. Selective dependency aggregation for action classification. *acm multimedia*.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2018. Learning to compose dynamic tree structures for visual contexts. *computer vision and pattern recognition*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need.

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

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *computer vision and pattern recognition*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *computer vision and pattern recognition*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Learning*.

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

Yue Wang, Shafiq Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C. H. Hoi. 2020. Vd-bert: A unified vision and dialog transformer with bert. *arXiv: Computer Vision and Pattern Recognition*.

Yuxuan Wang, Xueliang Zhao, and Dongyan Zhao. 2022b. NLPCC-2022-Shared-Task-4.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2022. Cpt: Colorful prompt tuning for pre-trained vision-language models.

Tong Ye, Shijing Si, Jianzong Wang, Rui Wang, Ning Cheng, and Jing Xiao. 2022. Vu-bert: A unified framework for visual dialog.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based localizing network for natural language video localization. *arXiv: Computation and Language*.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2019a. Learning 2d temporal adjacent networks for moment localization with natural language. *national conference on artificial intelligence*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *meeting of the association for computational linguistics*.

Zhuosheng Zhang and Hai Zhao. 2021. Advances in multi-turn dialogue comprehension: A survey. *arXiv: Computation and Language*.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, and Jie Tang. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv: Computation and Language*.