

# BI-MDRG: Bridging Image History in Multimodal Dialogue Response Generation

Hee Suk Yoon<sup>1\*</sup>, Eunseop Yoon<sup>1\*</sup>, Joshua Tian Jin Tee<sup>1\*</sup>, Kang Zhang<sup>1</sup>, Yu-Jung Heo<sup>2</sup>, Du-Seong Chang<sup>2</sup>, and Chang D. Yoo<sup>1\*\*</sup>

<sup>1</sup> Korea Advanced Institute of Science and Technology (KAIST)

<sup>2</sup> KT Corporation

{hskyoon, esyoon97, joshuateetj, zhangkang, cd\_yoo}@kaist.ac.kr  
{yj.heo, dschang}@kt.com

**Abstract.** *Multimodal Dialogue Response Generation (MDRG)* is a recently proposed task where the model needs to generate responses in texts, images, or a blend of both based on the dialogue context. Due to the lack of a large-scale dataset specifically for this task and the benefits of leveraging powerful pre-trained models, previous work relies on the text modality as an intermediary step for both the image input and output of the model rather than adopting an end-to-end approach. However, this approach can overlook crucial information about the image, hindering 1) image-grounded text response and 2) consistency of objects in the image response. In this paper, we propose **BI-MDRG** that bridges the response generation path such that the image history information is utilized for enhanced relevance of text responses to the image content and the consistency of objects in sequential image responses. Through extensive experiments on the multimodal dialogue benchmark dataset, we show that BI-MDRG can effectively increase the quality of multimodal dialogue. Additionally, recognizing the gap in benchmark datasets for evaluating the image consistency in multimodal dialogue, we have created a curated set of 300 dialogues annotated to track object consistency across conversations. The code and the dataset is publicly available at <https://github.com/hee-suk-yoon/BI-MDRG>.

**Keywords:** Multimodal Dialogue · Image Grounding · Image Consistency

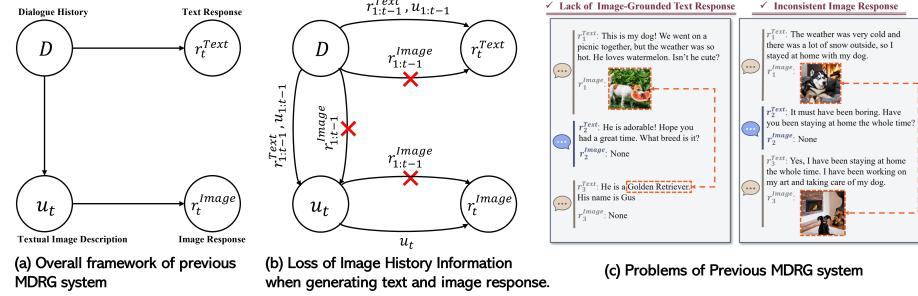
## 1 Introduction

With the development of instant messaging technology, visual modalities are increasingly used alongside text in online communication. To enhance user interaction with intelligent agents, a new task, Multimodal Dialogue Response Generation (MDRG) [43], has been proposed. This task requires models to generate both text and image responses based on dialogue history containing texts and images. Since learning an effective multimodal generation model with a sin-

---

\* Equal Contribution

\*\* Corresponding Author



**Fig. 1:** (a) Outlines the framework of previous Multimodal Dialogue Response Generation (MDRG) systems, which uses the textual descriptions of images ( $u_t$ ) as an intermediary step toward generating image responses ( $r_t^{Image}$ ). (b) Highlights the limitations of these systems, particularly their failure to fully leverage image history ( $r_{1:t-1}^{Image}$ ) in crafting both the textual response ( $r_t^{Text}$ ) and the image response ( $r_t^{Image}$ ). (c) Illustrates the consequences of this oversight, including responses that lack grounding in image context and consistency in image-based replies.

gle sequence-to-sequence model requires a large number of training instances, the lack of a large-scale multimodal dialogue dataset poses a significant challenge for the task.

To solve this issue, the previous MDRG model [43] disentangles the textual response and image response generation, as illustrated in Figure 1-(a). Specifically, given the dialogue history  $D$  at  $t$ -th conversation turn, the MDRG model first generates the textual response  $r_t^{Text}$  and the intermediary textual image description  $u_t$ . This image description  $u_t$  is subsequently fed into a text-to-image model to generate the corresponding image  $r_t^{Image}$ . Such an approach allows the utilization of powerful models pre-trained on vast amounts of available data for text-to-text and text-to-image pairs, bypassing the need for a large-scale multimodal dialogue dataset for direct end-to-end training.

Nonetheless, due to its reliance on text as an intermediary representation of images, the previous MDRG model overlooks crucial information about the image, as shown in Figure 1-(b). By converting images from the dialogue history into textual descriptions ( $u_{1:t-1}$ ), these models fail to fully utilize the rich visual content in the actual images ( $r_{1:t-1}^{Image}$ ). This leads to two major issues: a lack of image-grounded context in textual responses (Figure 1-(c) (left)) and inconsistencies in image generation across dialogues (Figure 1-(c) (right)).

For instance, Figure 1-(c) (left) shows the existing MDRG model failing to provide an image-grounded response to ‘What breed is your dog?’ because it only perceives the dog through the text ‘a dog eating a watermelon’. Figure 1-(c) (right) illustrates the model’s inconsistency, where the ‘dog’ in the image history is not maintained in the generated response. These issues highlight the need for improved MDRG models that effectively utilize image history.

**Contribution** This paper introduces BI-MDRG to enhance text and image responses by bridging the image history.

- *Bridging Image History in Text Responses:* In Section 3.1, we propose an architectural modification in which the visual features, extracted using a visual encoder, are integrated into the cross-attention layers of the core language model. This is complemented by a novel multimodal causal attention mask modulation tailored for the MDRG task to improve image-grounded textual responses.
- *Bridging Image History in Image Responses:* In Section 3.2, we propose a citation framework where we use the Citation Module designed to augment textual image descriptions by tagging objects with citation tags that identify recurring objects throughout a dialogue. In Section 3.3, we show that training with these augmented data enables the model to recognize and maintain the consistency of objects in subsequent image responses during inference using Customized Text-to-Image Models. Due to the absence of benchmark datasets for evaluating image consistency in multimodal dialogues, we have created the Multimodal Dialogue Image Consistency (MDIC) dataset containing dialogues annotated to track object consistency across conversations.

## 2 Related Work

### 2.1 Multimodal Dialogue Datasets

Multimodal dialogue datasets generally fall into three categories: question and answering [1, 8] (where the task involves asking and answering questions about a specific image), in-scene [12, 31, 35, 44, 49] (where each dialogue turn corresponds to a scene from a movie or video), and conversation-based [9, 20, 21, 27, 42, 47] (which engage in natural dialogue about a given image or involves image sharing within natural conversations). Further details are in Appendix C.1.

This paper primarily explores the model’s capability for natural dialogue within the conversation-based category. Notable datasets in this segment include ImageChat [42], PhotoChat [47], MMDD [20], DialogCC [21], MMChat [51], TikTalk [27], and MMDDialog [9]. Given our focus on English-language scenarios, we exclude MMChat and TikTalk from our evaluation, as it is a dataset primarily in Chinese. Additionally, DialogCC (not publicly available) and MMDD, which are synthesized by algorithmically pairing images with text-only dialogues for random turns, are also excluded from our analysis. Therefore, our evaluation is centered on ImageChat, PhotoChat, and MMDDialog.

ImageChat [42] consists of image-centered dialogues, where each dialogue is centered around a single given image. PhotoChat [47] features dialogues collected from social media, where a single image is shared in one of the conversation turns, which better mirrors everyday human interaction. Still, their limited scale and domain diversity restrict their applicability. Overcoming these limitations, MMDDialog [9] features over a million diverse dialogues from social media, where multiple images are shared across numerous conversation turns, providing a more realistic representation of open-domain multimodal conversations.

## 2.2 Multimodal Dialogue Modeling

Pioneering studies [11, 32, 36] have delved into improving the performance of image-grounded conversational agents. While [14, 45, 46, 51] introduced a Seq2Seq based model focusing on multimodal dialogues, it primarily generated textual responses, not fully embracing the multimodal response scenario. In a notable advancement, [43] presented Divter, which not only produces informative text but also generates high-resolution images, marking a significant leap forward in multimodal dialogue response generation (MDRG). *It is important to note that our focus in this paper is generation-based models, contrasting with retrieval-based [3, 16, 22], which output image responses by retrieving existing images from a corpus instead of generating new ones.*

## 2.3 Customized Text-to-Image

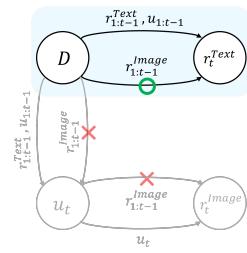
Recent studies on text-to-image diffusion models [5, 17, 38, 40] focus on customization [7, 10, 18, 39], learning specific concepts from a few images. Following this, users can flexibly generate the learned concepts into new scenes. Text inversion [10] generates varied contexts for a concept by updating the text embedding without altering the model. Dreambooth [39] and Custom Diffusion [18] fine-tune the U-Net architecture using an identifier, class label, and images. In a notable enhancement, BLIP-Diffusion [23] enables zero-shot subject-driven generation, allowing fast customized text-to-image generation.

## 3 BI-MDRG: Bridging the Image History in Multimodal Dialogue Response Generation

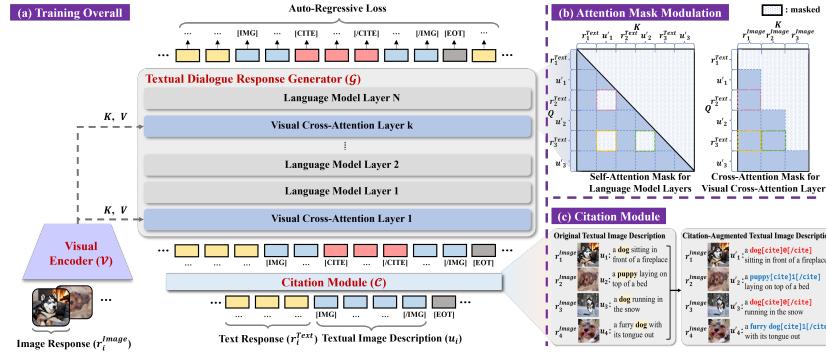
We introduce BI-MDRG, a conversational agent designed to produce both textual and visual responses with enhanced awareness of image history. Sections 3.1 and 3.2 detail the training procedure, effectively integrating image history information into text responses and textual image descriptions. Section 3.3 outlines the inference process, wherein the image history informs the image responses by leveraging the captured details from enhanced textual image descriptions.

### 3.1 Bridging the Image History for Image-Grounded Text Response

The example shown in Figure 1-(c) (*left*) highlighted one of the crucial limitations in the previous MDRG system: their reliance on textual descriptions for understanding image history, which hinders the image-grounded textual responses. To overcome this, we have adopted an architectural change along with a multimodal causal attention mask modulation to effectively bridge the image history information  $r_{1:t-1}^{Image}$  to the text response  $r_t^{Text}$  (as depicted in Figure 2).



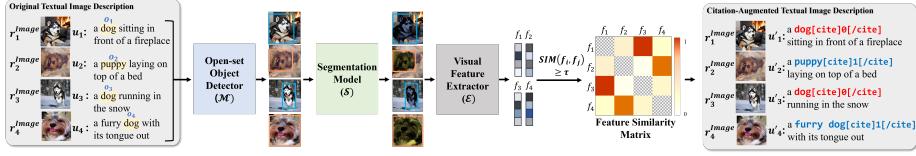
**Fig. 2:** Bridging Image History to the Text Response.



**Fig. 3: Training of BI-MDRG.** (a) Textual Dialogue Response Generator  $\mathcal{G}$  cross-attends to the image features from the Visual Encoder  $\mathcal{V}$ . (b) Attention Mask Modulation alters the causal attention to prioritize image features over textual image descriptions. (c) Citation Module  $\mathcal{C}$  generates citation-augmented textual image descriptions, enabling the tracking of objects within image history for consistency maintenance.

**Architecture** Our Textual Dialogue Response Generator  $\mathcal{G}$  (Figure 3-(a)), consists of a decoder-only language model with added visual cross-attention layers. These layers directly engage with image features provided by the Visual Encoder  $\mathcal{V}$ , drawing inspiration from Flamingo [2], to reduce dependence on textual image descriptions for perceiving images. In our framework, a dialogue context  $D = \{(r_i^{\text{Text}}, r_i^{\text{Image}})\}_{i=1}^n$  comprises multiple turns, each with an associated text response  $r_i^{\text{Text}}$  and an image response  $r_i^{\text{Image}}$ . An image captioning model produces textual descriptions  $u_i$  for each image  $r_i^{\text{Image}}$ , which are then transformed into citation-augmented descriptions  $u'_i$  as shown in Figure 3-(c) and further detailed in Section 3.2 (*if there is no image response for a turn,  $r_i^{\text{Image}} = \emptyset$  and  $u'_i = \emptyset$* ). These descriptions and the text responses  $\{r_1^{\text{Text}}, u'_1, \dots, r_n^{\text{Text}}, u'_n\}$  are fed into  $\mathcal{G}$ , while the images  $\{r_1^{\text{Image}}, \dots, r_n^{\text{Image}}\}$  are processed by  $\mathcal{V}$  to extract image features which are fed to the cross-attention layers in  $\mathcal{G}$ . Although our model directly cross-attends to the inputted image features, we retain the textual descriptions  $u'_{1:n}$  as essential inputs since we require the generation of textual description by our  $\mathcal{G}$  model, which subsequently gets used by a text-to-image model for constructing the image response. Keeping the textual description inputs allows efficient teacher-forced next token prediction training.

**Multimodal Causal Attention Mask Modulation** For the input sequence  $\{r_1^{\text{Text}}, u'_1, \dots, r_n^{\text{Text}}, u'_n\}$ , we use a specialized mask along with the standard causal mask (Figure 3-(b)). The traditional causal mask allows each text response  $r_i^{\text{Text}}$  to access previous textual image descriptions  $u'_{1:i-1}$ , leading to reliance on textual information over visual context. Our masking strategy prevents  $r_i^{\text{Text}}$  from accessing  $u'_{1:i-1}$ , redirecting focus to the actual image features of  $r_{1:i-1}^{\text{Image}}$ , ensuring text responses are grounded on raw image features.



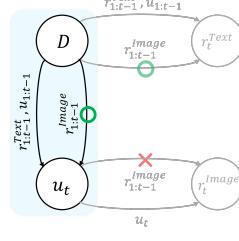
**Fig. 4: Illustration of the Citation Module.** Citation Module recognizes identical objects within image history and injects this information into the textual image description with citation tags (e.g., [cite]0[/cite]).

### 3.2 Citation Module: Bridging the Image History to the Textual Image Description

The example shown in Figure 1-(c) (*right*) highlighted another critical limitation of the previous MDRG system: its inability to ensure consistency in image responses. To address this, we propose the Citation Module (Figure 3-(c)) to bridge the image history  $r_{1:t-1}^{\text{Image}}$  and the textual image description  $u_t$  (as depicted in Figure 5) by ensuring that the textual image description accurately relays which objects should persist in subsequent images.

**Citation Module** Citation Module plays a pivotal role in tracking recurring objects in the dialogue using textual image descriptions. For instance, descriptions like "a dog is in front of a fireplace" and "a dog running in the snow" are augmented to "a dog[cite]0[/cite] is in front of a fireplace" and "a dog[cite]0[/cite] running in the snow," respectively if they reference the same dog.

Motivated by [30], Figure 4 details the citation process for textual image descriptions  $\{u_1, \dots, u_n\}$ . For each textual image description  $u_i$ , a Part of Speech (POS) tagging processor  $\mathcal{P}$  is employed to tag words and pinpoint the key object word  $o_i$  in the description. The word  $o_i$ , along with its corresponding image  $r_i^{\text{Image}}$ , is processed through an open-set object detector  $\mathcal{M}$  to obtain the bounding box of the detected object, which is then input to a segmentation model  $\mathcal{S}$  for generating object segmentation mask  $s_i$ . These masks are applied to isolate the objects from their backgrounds in  $r_i^{\text{Image}}$  since background removal has been proven helpful for better extraction of object features [6]. These isolated objects are then analyzed by a visual feature extractor  $\mathcal{E}$  to extract features  $f_i$ . The resulting feature set  $\{f_1, \dots, f_n\}$  undergoes clustering based on cosine similarity to identify identical objects across images, as outlined in Algorithm 1; this involves assigning each element a cluster id  $c_i$  based on the similarity of their features. For each  $o_i$  in  $u_i$ , we augment the word so that it is followed by its corresponding cluster id (i.e., citation tag)  $c_i$  of  $f_i$ , resulting in the citation augmented textual description  $u'_i$ . This Citation Module operates with off-the-shelf components for citation tags, requiring no training on the target dataset.



**Fig. 5:** Bridging Image History to the Textual Image Description.

**Algorithm 1** Citation Module

---

**Require:** image response  $(r_1^{\text{Image}}, \dots, r_n^{\text{Image}})$ , textual image description  $(u_1, \dots, u_n)$ , similarity threshold  $\tau$ , POS tagging processor  $\mathcal{P}$ , open-set object detector  $\mathcal{M}$ , segmentation model  $\mathcal{S}$ , visual feature extractor  $\mathcal{E}$ , object cluster dictionary  $\mathcal{K} = \{\}$

```

for  $i \leftarrow 1$  to  $n$  do
     $o_i \leftarrow \mathcal{P}(u_i)$                                  $\triangleright$  Identify principal object word in  $u_i$ .
     $box_i \leftarrow \mathcal{M}(o_i, r_i^{\text{Image}})$            $\triangleright$  Obtain bounding box for  $o_i$ .
     $s_i \leftarrow \mathcal{S}(box_i)$                           $\triangleright$  Obtain segmentation mask for  $o_i$ .
     $f_i \leftarrow \mathcal{V}(s_i \odot r_i^{\text{Image}})$             $\triangleright$  Extract visual feature for  $o_i$ .
end for

 $cluster\_id \leftarrow 0$ 
for  $i \leftarrow 1$  to  $n$  do                                 $\triangleright$  Perform clustering of  $o_i$ .
    if not  $\mathcal{K}.\text{has\_key}(o_i)$  then
         $\mathcal{K}[o_i] \leftarrow cluster\_id$                    $\triangleright$  Gets own cluster.
         $cluster\_id \leftarrow cluster\_id + 1$ 
    end if
    for  $j \leftarrow i + 1$  to  $n$  do
        if  $\text{SIM}(f_i, f_j) \geq \tau$  and not  $\mathcal{K}.\text{has\_key}(o_j)$  then
             $\mathcal{K}[o_j] \leftarrow \mathcal{K}[o_i]$                    $\triangleright$  Puts into existing cluster.
        end if
    end for
end for
return  $\mathcal{K}$ 

```

---

**Generative Training Objective** We use the next token prediction training via teacher forcing, which is used in standard auto-regressive language models [37]. Specifically, given the token sequence  $w = \{w_j\}_{j=1}^N$  of the input sequence  $\{r_1^{\text{Text}}, u'_1, \dots, r_n^{\text{Text}}, u'_n\}$  and the images  $\{r_1^{\text{Image}}, \dots, r_n^{\text{Image}}\}$ , we minimize the negative log-likelihood:

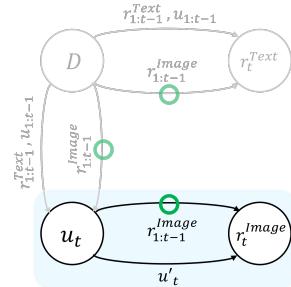
$$\mathcal{L}(w) = \sum_{j=1}^N \log P(w_j | w_{<j}, \{r_i^{\text{Image}}\}_{i=1}^n; \mathcal{G}, \mathcal{V}). \quad (1)$$

With such training, our model can generate textual image descriptions during inference with citation tags that reflect the objects needing consistency.

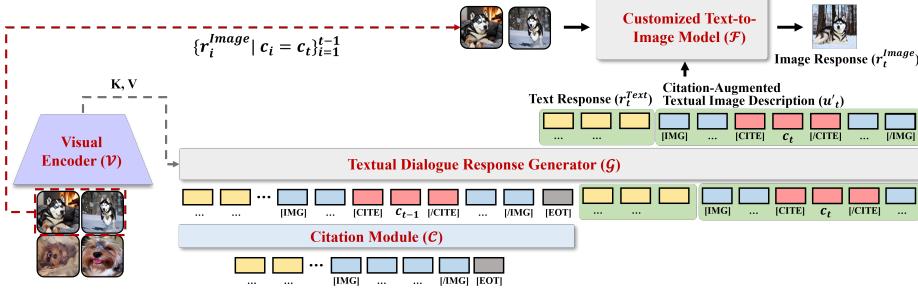
### 3.3 Inference Procedure: Bridging the Image History for Consistent Image Response

This section outlines our inference procedure, which employs Customized Text-to-Image Model [23] in conjunction with citation-augmented textual image descriptions (Figure 7). This bridges the image history  $r_{1:t-1}^{\text{Image}}$  and image response  $r_t^{\text{Image}}$  (as depicted in Figure 6) allowing for consistent generation of the image response.

**Inference** For an incoming dialogue context  $D = \{(r_i^{\text{Text}}, r_i^{\text{Image}})\}_{i=1}^{t-1}$ , we initially construct



**Fig. 6:** Bridging the Image History to the Image Response.



**Fig. 7: Overall Inference Procedure.** During inference, the Textual Response Generator  $\mathcal{G}$  generates the citation-augmented textual image description  $u'_t$ . By utilizing its citation tag  $c_t$ , object consistency can be maintained by feeding all preceding image responses with the identical citation tag into a Customized Text-to-Image Model  $\mathcal{F}$ .

the sequence of textual image descriptions and corresponding text responses as  $\{r_1^{\text{Text}}, u_1, \dots, r_{t-1}^{\text{Text}}, u_{t-1}\}$ . Our Citation Module  $\mathcal{C}$  then augments the textual image descriptions with citation tags to create  $\{r_1^{\text{Text}}, u'_1, \dots, r_{t-1}^{\text{Text}}, u'_{t-1}\}$ , which serves as the input for the Textual Response Generator  $\mathcal{G}$ . During inference, if  $\mathcal{G}$  predicts an [IMG] token, it initiates the generation of a citation-augmented textual image description  $u'_t$ . We extract the citation  $c_t$  and the core image description  $u_t$  from  $u'_t$ . Subsequently, the image description  $u_t$  is input to our Customized Text-to-Image Model  $\mathcal{F}$ , along with all preceding image responses sharing the same citation  $c_t$  as follows:

$$r_t^{\text{Image}} = \mathcal{F}(u_t | \{r_i^{\text{Image}} | c_i = c_t\}_{i=1}^{t-1}). \quad (2)$$

This approach ensures the consistent generation of specific objects across the conversation.

## 4 Experiments

This section presents the implementation details, evaluation benchmarks, and the experimental results of our approach. Section 4.1 outlines the implementation specifics. In Section 4.2, we assess the overall quality of our model against the standard benchmarks established in previous works [9, 47]. Section 4.3 is dedicated to evaluating the image grounding effectiveness of our model. Finally, Section 4.4 examines the consistency of the image responses generated by our system.

### 4.1 Experimental Setup

**Dataset** As mentioned in Section 2.1, our benchmark datasets include ImageChat [42], PhotoChat [47], and MMDialog [9]. Specifically, for the overall

multimodal dialogue evaluation in Section 4.2, we train and evaluate using the PhotoChat and MMDialog. For the image grounding evaluation in Section 4.3, we use ImageChat. For the image consistency evaluation in Section 4.4, we create the MDIC dataset by hand-labeling a subset of MMDialog test set. For Section 4.3 and Section 4.4, we use the model trained on MMDialog from Section 4.2.

**Implementation Details** For the Textual Dialogue Response Generator  $\mathcal{G}$  and the Visual Encoder  $\mathcal{V}$ , we initialize with the pre-trained OpenFlamingo 4B model<sup>3</sup> [4]. During the fine-tuning phase, we employ special tokens to structure our inputs and outputs: `[IMG]` and `[/IMG]` encapsulate textual image descriptions, while `[EOT]` signifies the end of conversation turns. The BLIP2-flan-t5-xl model [24] is used for converting image responses to corresponding textual image descriptions. Additionally, our Citation Module  $\mathcal{C}$  uses `[CITE]` and `[/CITE]` tokens to mark the beginning and end of citations linked to key objects within the textual image descriptions.

As noted in Section 3.2, the Citation Module  $\mathcal{C}$  is composed of four key components: the POS tagging processor ( $\mathcal{P}$ ), the open-set object detector ( $\mathcal{M}$ ), the segmentation model ( $\mathcal{S}$ ), and the visual feature extractor ( $\mathcal{E}$ ). Specifically, we employ spaCy [13] for  $\mathcal{P}$ , GroundingDino [28] for  $\mathcal{M}$ , the Segment Anything Model (SAM) [15] for  $\mathcal{S}$ , and DINOv2 [33] for  $\mathcal{E}$ . The similarity threshold  $\tau$  is set to 0.6. No further learning is done for these modules; they are utilized as pre-trained components within our system for citation tagging. For our customized text-to-image generation model,  $\mathcal{F}$ , we used BLIP-Diffusion [23] when conditioning on the input image and the standard Stable Diffusion 2.1 [38] when there is no input image conditioning.

**Learning Details** Let us denote  $\{\theta_{\mathcal{V}}, \theta_{\mathcal{G}_v}, \theta_{\mathcal{G}_l}\}$  as the parameters of the perceiver resampler of the Visual Encoder  $\mathcal{V}$ , the visual cross-attention layers of the Textual Dialogue Response Generator  $\mathcal{G}$ , and the language model layers of  $\mathcal{G}$ , respectively.

In the first stage of training, we train  $\theta_{\mathcal{G}_l}$ . The batch size is set to 256 with a maximum token length set to 256. In the second stage of training, we jointly train  $\theta_{\mathcal{V}}$  and  $\theta_{\mathcal{G}_v}$ . The batch size is set to 128 with a maximum token length set to 512. Both the first and second stage is trained by minimizing the next token prediction loss (Eq. 1) using the AdamW optimizer [29] with a learning rate set to 1e-4. The trainings were conducted using 16 x NVIDIA A100 80GB PCIe.

## 4.2 Multimodal Dialogue Evaluation

**Evaluation Dataset** We evaluate the overall performance of our BI-MDRG system on the test set of PhotoChat [47] and MMDialog [9] dataset. PhotoChat contains a single image per dialogue, while the MMDialog includes dialogues with multiple images across turns, offering a more complex context for assessing

---

<sup>3</sup> <https://huggingface.co/openflamingo/OpenFlamingo-4B-vitl-rpj3b-langinstruct>

**Table 1:** Automatic evaluation results of BI-MDRG on the PhotoChat and the MMDialog test set. Numbers in bold represent the best scores.

Model	Intent	Image	Response	Textual		Image Description		Text Response			
	F1	IS		B1	B2	R-1	R-L	B1	B2	R1	R-L
<b>PhotoChat Dataset [47]</b>											
BERT-base [47]	53.2	-	-	-	-	-	-	-	-	-	-
T5-3B [47]	<b>58.9</b>	-	-	-	-	-	-	-	-	-	-
Divter [43]	56.2	15.8	15.1	11.4	-	15.8	6.52	1.66	-	5.69	
Divter <sub>LLM</sub>	54.1	16.1	41.3	27.1	43.3	41.6	11.4	4.75	11.2	10.8	
<b>BI-MDRG</b>	55.7	<b>16.7</b>	<b>42.1</b>	<b>28.2</b>	<b>44.6</b>	<b>42.5</b>	<b>12.4</b>	<b>5.12</b>	<b>12.1</b>	<b>11.2</b>	
<b>MMDialog Dataset [9]</b>											
PaCE [25]	<b>77.6</b>	-	-	-	-	-	-	-	-	-	-
Divter [9]	71.8	20.5	-	-	-	-	9.44	7.45	-	11.2	
MiniGPT5 [50]	-	20.2	-	-	-	-	<b>29.1</b>	19.5	-	12.1	
Divter <sub>LLM</sub>	67.3	21.0	44.2	35.7	45.5	43.6	21.3	16.2	20.4	19.4	
<b>BI-MDRG</b>	70.5	<b>22.4</b>	<b>52.2</b>	<b>44.7</b>	<b>53.2</b>	<b>51.6</b>	27.6	<b>23.5</b>	<b>25.7</b>	<b>24.8</b>	

our system’s capability. We perform the tests for all turns except the first turn of each dialogue and consider all previous turns as context.

**Evaluation Metric** The performance evaluation is carried out using automatic metrics across four key dimensions: (1) Image Intent Prediction - assessing the need for an image response in the current turn; (2) Image Response Quality; (3) Textual Image Description Quality; and (4) Text Response Generation. We employ the F1 metric for Image Intent Prediction, following the binary classification approach detailed in [9,43]. BLEU [34] and ROUGE [26] metrics are used for evaluating both the Textual Image Descriptions and Text Responses. The Image Response Quality is measured using the Inception Score (IS) [41], in line with [9].

**Baselines** The BI-MDRG system is compared against Divter [43], a prior Multimodal Dialogue Response Generator that utilizes DialoGPT (762M) [48] as its language model backbone. For a fair comparison, we also train Divter on the same backbone Language Model as our BI-MDRG, denoted as Divter<sub>LLM</sub> (3B). Moreover, for the MMDialog dataset, we compare with MiniGPT5 [50] which uses a 9B backbone VLM. For Intent-prediction baselines, we also include BERT-base, T5-3B reported in [47] for the PhotoChat dataset and PaCE reported in [25] for the MMDialog dataset.

**Results** Table 1 presents the evaluation results, where our BI-MDRG system demonstrates outstanding performance on both datasets. In Textual Image Description and Text Response Generation, BI-MDRG achieves the highest BLEU and ROUGE metrics scores in most cases, indicating the system’s proficiency in generating relevant and coherent image responses and text responses. Notably,

BI-MDRG, which uses a 4B backbone VLM, outperforms MiniGPT5, which utilizes a 9B backbone VLM.

### 4.3 Image Grounding Evaluation

**Evaluation Dataset** To evaluate the ability to generate image-grounded textual responses, we utilize the ImageChat dataset [42], consisting of dialogues centered around a single image (an example is shown in Appendix C.4). Since the conversations are grounded in an image, the dataset is suitable for the evaluation of the image grounding capability of the model. We use the same model trained on the MMDialog dataset in Section 4.2 without further tuning on the ImageChat train set.

**Evaluation Metric** In order to evaluate the image grounding capability, we use BLEU and ROUGE as the metrics. We perform the evaluation only on the last turn of each dialogue and consider all previous turns as the input context.

**Results** The performance of our BI-MDRG system on the ImageChat benchmark is summarized in Table 2. These results demonstrate the system’s superior image grounding capabilities in text responses. BI-MDRG achieves the highest scores in BLEU-1, ROUGE-1, and ROUGE-L metrics, indicating its effectiveness in generating contextually relevant and coherent text responses grounded in the visual elements of the conversation. In comparison to Divter<sub>LLM</sub> and BI-MDRG<sub>w/o mask</sub>, our BI-MDRG model shows an improvement, highlighting the significance of our proposed architectural enhancements and the multimodal causal attention mask in understanding and integrating visual context into text responses effectively.

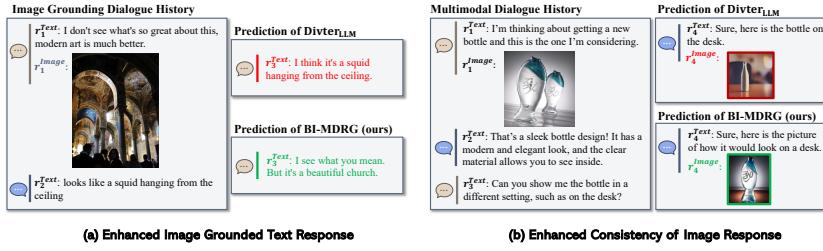
Moreover, we show an example of the model prediction between the Divter<sub>LLM</sub> and our BI-MDRG in Figure 8-(a). Figure 8-(a) (*right-top*) shows that Divter<sub>LLM</sub> produces contextually irrelevant text response due to its lack of image understanding. However, Figure 8-(a) (*right-bottom*) shows that since our BI-MDRG has access to the image, it can produce a text response grounded on the image.

### 4.4 Image Response Consistency Evaluation

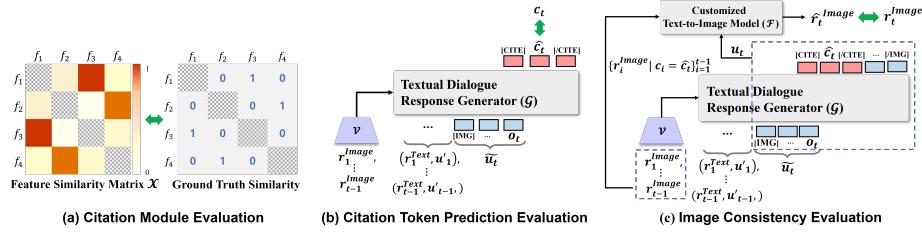
We integrated the Citation Module into our BI-MDRG system to improve consistency in image responses within the dialogue. This module assigns citation tags to primary objects in textual image descriptions, tracking their presence across

**Table 2:** Automatic evaluation results on the ImageChat test set. Numbers in bold represent the best scores.

Model	B1	R-1	R-L
<b>ImageChat Dataset [42]</b>			
Divter <sub>LLM</sub>	8.6	10.3	9.6
BI-MDRG <sub>w/o mask</sub>	10.0	11.1	10.2
<b>BI-MDRG</b>	<b>10.9</b>	<b>11.7</b>	<b>10.9</b>



**Fig. 8:** Example of predictions from DivterLLM [43] vs. BI-MDRG (ours). Additional examples can be found in Appendix E.



**Fig. 9: Three evaluation aspects for image response consistency.** (a) We assess the accuracy of the pseudo-labels of citation tags created by the Citation Module. (b) We evaluate how well the model learns from these pseudo-labels and predict the citation tag  $\hat{c}_t$ , given the dialogue history, current text response  $r_t^{\text{Text}}$ , and  $\tilde{u}_t$  which denotes the textual image description leading up to the primary object word  $o_t$ . (c) We evaluate the consistency of the resulting image response  $\hat{r}_t^{\text{Image}}$ .

the dialogue. As shown in Figure 9, we focus on three aspects for evaluating the image response consistency: (a) *Citation Module Evaluation*, (b) *Citation Token Prediction Evaluation*, and (c) *Image Consistency Evaluation*.

**Multimodal Dialogue Image Consistency (MDIC) Dataset Creation**  
Due to the absence of benchmark datasets for evaluating the image consistency in multimodal dialogue scenarios, we created a dataset of 300 dialogues, each manually annotated to assign correct citation tags to objects based on their appearances in image responses. We use this created dataset for the subsequent evaluations. Details of the annotation process can be found in Appendix C.5.

**Citation Module Evaluation** Figure 9-(a) shows the evaluation method for the Citation Module. As described in Section 3.2, the module  $\mathcal{C}$  processes visual features  $\{f_1, \dots, f_n\}$  corresponding to primary objects  $\{o_1, \dots, o_n\}$  in image descriptions. The module generates citations based on the cosine similarity between these features. We constructed a similarity matrix  $\mathcal{X}$ , where each element  $\mathcal{X}_{ij}$  represents the cosine similarity between the features  $f_i$  and  $f_j$ . A feature pair  $(f_i, f_j)$  is deemed ‘similar’ if  $\mathcal{X}_{ij} \geq \tau$  and ‘not similar’ otherwise,

where  $\tau$  is a predetermined threshold. The module’s performance was quantified using the F1 score, and we achieved an F1 score of 0.72, demonstrating the module’s high accuracy in assigning pseudo-labels of citation tags for training.

### Citation Token Prediction Evaluation

Our model, trained with citation-augmented textual image descriptions, can generate these descriptions with accurate citation tags for the primary objects in the descriptions. Figure 9-(b) shows the evaluation method where we assess our model’s ability to predict the citation tags correctly. We input the ground truth sequence  $\{r_1^{\text{Text}}, u'_1, \dots, r_t^{\text{Text}}, \tilde{u}_t\}$  from the MDIC dataset along with the image response history  $\{r_1^{\text{Image}}, \dots, r_{t-1}^{\text{Image}}\}$ . Here,  $\tilde{u}_t$  denotes the textual image description leading up to the principal object word  $o_t$ . The goal is to verify whether the model’s predicted citation  $\hat{c}_t$  aligns with the actual citation  $c_t$ . The accuracy (Acc.) of our citation token prediction is presented in Table 3. We compare with a baseline denoted as Divter<sub>LLM</sub> + LLMCite where we assign citation tags by instructing an external LLM (Details in Appendix D). The results show that BI-MDRG achieves 84.0% accuracy compared to 33.5% of the LLMCite baseline, showing the effectiveness of our model in identifying objects needing consistency.

**Image Consistency Evaluation** Figure 9-(c) shows the evaluation method for the Image Consistency Evaluation. To assess the consistency of the image response, we employ metrics commonly found in the literature on custom object text-to-image generation [7, 10, 18, 19, 39]. DINO features outperform CLIP in capturing image alignment, specifically in distinguishing different objects of the same class [39]. Therefore, for our evaluation, we measure image alignment scores using the DINOv2 [33] features.

In our proposed framework, correctly using the citation tag is essential to ensure image consistency in the conversation. This involves correctly predicting a new citation tag or referencing an existing tag in history. Therefore, our evaluation includes not only scenarios where the model’s predicted citation tag aligns with the ground truth ( $c_t = \hat{c}_t$ ) and  $c_t$  exists in previous dialogue history, but also instances where the model’s predicted citation tag do not match the ground truth ( $c_t \neq \hat{c}_t$ ) and either  $\hat{c}_t$  or  $c_t$  is in previous dialogue history.

We report the DINOv2 score as an average of both the successful and failed citation cases mentioned above, where successful prediction positively influences the score, and failed prediction reduces it, offering a comprehensive measure of the model’s performance in image consistency. Table 3 shows that our method successfully maintains the consistency of the image response in a dialogue setting, with the DINOv2 score achieving 0.53 compared to 0.32 of baseline. Furthermore,

**Table 3:** Citation Token Prediction (Acc.) and Image Consistency Evaluation (DINOv2). Numbers in bold represent the best scores.

Model	Acc.	DINOv2( $\uparrow$ )
<b>MDIC Dataset</b>		
Divter <sub>LLM</sub> + LLMCite	33.5	0.32
<b>BI-MDRG (ours)</b>	<b>84.0</b>	<b>0.53</b>

**Table 4: Evaluation of the citation framework on various settings.** The image consistency is measured using DINOv2 ( $\uparrow$ ) on MDIC dataset, and TID (Textual Image Description) and TR (Text Response) are measured on MMDialog [9] test set. Custom uses customized text-to-image generation with image conditioning, while Diffusion employs a standard text-to-image diffusion model without conditioning. The citation framework allows to selectively use between these for consistent image response. *The first row represents the performance of BI-MDRG (ours).*

Citation	VLM Size	Diffusion	DINOv2	TID				TR			
				B1	B2	R-1	R-L	B1	B2	R-1	R-L
<b>Citation Module</b>	4B	Custom+Diffusion	<b>0.53</b>	52.2	44.7	53.2	51.6	27.6	23.5	25.7	24.8
LLMCite x	4B	Custom+Diffusion	0.34	52.0	44.9	53.3	51.3	27.4	23.4	25.8	24.6
	4B	Diffusion	0.25	52.0	44.9	53.3	51.3	27.4	23.4	25.8	24.6
LLMCite x	9B	Custom+Diffusion	0.33	<b>56.0</b>	<b>48.7</b>	<b>54.5</b>	<b>52.1</b>	<b>30.3</b>	<b>25.7</b>	<b>28.8</b>	<b>26.8</b>
	9B	Diffusion	0.26	<b>56.0</b>	<b>48.7</b>	<b>54.5</b>	<b>52.1</b>	<b>30.3</b>	<b>25.7</b>	<b>28.8</b>	<b>26.8</b>

in Figure 8-(b), we present a comparative example of enhanced image consistency of prediction from **BI-MDRG** compared to **Divter<sub>LLM</sub>**.

## 5 Importance of Citation Tags for Image Consistency

Due to powerful pre-trained models for text-to-text and text-to-image, adopting text as an intermediary for image responses is a practical solution for the MDRG [43] task. However, due to the inherent information loss of images during this process, achieving *image consistency is infeasible without a targeted framework for consistency maintenance*.

Table 4 shows image consistency and dialogue response performance across various settings. Not using citations shows similar dialogue response performance (Intent, TID, TR) compared to LLMCite and our Citation Module. However, LLMCite improves image consistency (DINOv2) from 0.25 to 0.34 (4B) and 0.26 to 0.33 (9B). Our Citation Module further boosts this from 0.34 to 0.53, indicating its importance for image consistency without affecting dialogue response performance. Notably, scaling the model size (4B to 9B) improves textual response but fails to maintain image consistency without our citation framework, as also evident by ChatGPT’s shortcomings in Appendix E.

## 6 Conclusion

This paper presents BI-MDRG, a novel framework for Multimodal Dialogue Response Generation (MDRG) aimed at bridging the image history for enhanced text and image response. Our model’s innovative use of image history to inform both text and image responses addresses fundamental limitations in previous methodologies, particularly in maintaining consistency in multimodal interactions. The effectiveness of BI-MDRG has been demonstrated through rigorous evaluations using multiple benchmark datasets and a custom-annotated dataset.

## Acknowledgement

This work was supported by a grant of the KAIST-KT joint research project through AI2X Lab., Tech Innovation Group, funded by KT (No. D23000019, Developing Visual and Language Capabilities for AI-Based Dialogue Systems), and by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-01381, Development of Causal AI through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments).

## References

1. Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., Lee, S., Parikh, D.: Audio-visual scene-aware dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), <https://openreview.net/forum?id=EbMuimAbPbs>
3. Anonymous: Chatsearch: a dataset and a generative retrieval model for general conversational image retrieval (2023), <https://openreview.net/forum?id=0unbjYPmbC>
4. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
5. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., Karras, T., Liu, M.Y.: ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
6. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)
7. Cohen, N., Gal, R., Meirom, E.A., Chechik, G., Atzmon, Y.: “this is my unicorn, fluffy”: Personalizing frozen vision-language representations. In: European Conference on Computer Vision. pp. 558–577. Springer (2022)
8. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 326–335 (2017)
9. Feng, J., Sun, Q., Xu, C., Zhao, P., Yang, Y., Tao, C., Zhao, D., Lin, Q.: MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7348–7363. Association for Computational Linguistics, Toronto, Canada (Jul 2023). <https://doi.org/10.18653/v1/2023.acl-long.405>, <https://aclanthology.org/2023.acl-long.405>

10. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022). <https://doi.org/10.48550/ARXIV.2208.01618>, <https://arxiv.org/abs/2208.01618>
11. Gan, Z., Cheng, Y., Kholy, A., Li, L., Liu, J., Gao, J.: Multi-step reasoning via recurrent dual attention for visual dialog. In: Korhonen, A., Traum, D., Márquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6463–6474. Association for Computational Linguistics, Florence, Italy (Jul 2019). <https://doi.org/10.18653/v1/P19-1648>, <https://aclanthology.org/P19-1648>
12. Han, S., Hessel, J., Dziri, N., Choi, Y., Yu, Y.: Champagne: Learning real-world conversation from large-scale web videos. arXiv preprint arXiv:2303.09713 (2023)
13. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
14. Kim, J., Yoon, S., Kim, D., Yoo, C.D.: Structured co-reference graph attention for video-grounded dialogue (2021), <https://arxiv.org/abs/2103.13361>
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
16. Koh, J.Y., Salakhutdinov, R., Fried, D.: Grounding language models to images for multimodal inputs and outputs. ICML (2023)
17. Koo, G., Yoon, S., Yoo, C.D.: Wavelet-guided acceleration of text inversion in diffusion-based image editing. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4380–4384. IEEE (2024)
18. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion (2023)
19. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion (2023)
20. Lee, N., Shin, S., Choo, J., Choi, H.J., Myaeng, S.H.: Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 897–906. Association for Computational Linguistics, Online (Aug 2021), <https://aclanthology.org/2021.acl-short.113>
21. Lee, Y.J., Ko, B., Kim, H.G., Choi, H.J.: Dialogcc: Large-scale multi-modal dialogue dataset. arXiv preprint arXiv:2212.04119 (2022)
22. Levy, M., Ben-Ari, R., Darshan, N., Lischinski, D.: Chatting makes perfect: Chat-based image retrieval. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 61437–61449. Curran Associates, Inc. (2023), [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/c1b3d1e2cf53bb28cabd801bd58b3521-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/c1b3d1e2cf53bb28cabd801bd58b3521-Paper-Conference.pdf)
23. Li, D., Li, J., Hoi, S.C.H.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing (2023)
24. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)

25. Li, Y., Hui, B., Yin, Z., Yang, M., Huang, F., Li, Y.: PaCE: Unified multi-modal dialogue pre-training with progressive and compositional experts. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13402–13416. Association for Computational Linguistics, Toronto, Canada (Jul 2023). <https://doi.org/10.18653/v1/2023.acl-long.749>, <https://aclanthology.org/2023.acl-long.749>
26. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), <https://aclanthology.org/W04-1013>
27. Lin, H., Ruan, L., Xia, W., Liu, P., Wen, J., Xu, Y., Hu, D., Song, R., Zhao, W.X., Jin, Q., et al.: Tiktalk: A video-based dialogue dataset for multi-modal chitchat in real world. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1303–1313 (2023)
28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations
30. Ma, J., Liang, J., Chen, C., Lu, H.: Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. arXiv preprint arXiv:2307.11410 (2023)
31. Meng, Y., Wang, S., Han, Q., Sun, X., Wu, F., Yan, R., Li, J.: Opennvidial: A large-scale, open-domain dialogue dataset with visual contexts. arXiv preprint arXiv:2012.15015 (2020)
32. Niu, Y., Zhang, H., Zhang, M., Zhang, J., Lu, Z., Wen, J.R.: Recursive visual attention in visual dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6679–6688 (2019)
33. Oquab, M., Dariseti, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
34. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. p. 311–318. ACL '02, Association for Computational Linguistics, USA (2002). <https://doi.org/10.3115/1073083.1073135>, <https://doi.org/10.3115/1073083.1073135>
35. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: Korhonen, A., Traum, D., Márquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 527–536. Association for Computational Linguistics, Florence, Italy (Jul 2019). <https://doi.org/10.18653/v1/P19-1050>, <https://aclanthology.org/P19-1050>
36. Qi, J., Niu, Y., Huang, J., Zhang, H.: Two causal principles for improving visual dialog. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10860–10869 (2020)
37. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)

38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
39. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
40. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Gontijo-Lopes, R., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), <https://openreview.net/forum?id=08Ykn512A1>
41. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016), [https://proceedings.neurips.cc/paper\\_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf)
42. Shuster, K., Humeau, S., Bordes, A., Weston, J.: Image-chat: Engaging grounded conversations. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2414–2429. Association for Computational Linguistics, Online (Jul 2020). <https://doi.org/10.18653/v1/2020.acl-main.219>, <https://aclanthology.org/2020.acl-main.219>
43. Sun, Q., Wang, Y., Xu, C., Zheng, K., Yang, Y., Hu, H., Xu, F., Zhang, J., Geng, X., Jiang, D.: Multimodal dialogue response generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2854–2866. Association for Computational Linguistics, Dublin, Ireland (May 2022). <https://doi.org/10.18653/v1/2022.acl-long.204>, <https://aclanthology.org/2022.acl-long.204>
44. Wang, S., Meng, Y., Li, X., Sun, X., Ouyang, R., Li, J.: Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. arXiv preprint arXiv:2109.12761 (2021)
45. Yoon, S., Kim, D., Yoon, E., Yoon, H.S., Kim, J., Yoo, C.D.: Hear: Hearing enhanced audio response for video-grounded dialogue. arXiv preprint arXiv:2312.09736 (2023)
46. Yoon, S., Yoon, E., Yoon, H.S., Kim, J., Yoo, C.: Information-theoretic text hallucination reduction for video-grounded dialogue. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 4182–4193. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). <https://doi.org/10.18653/v1/2022.emnlp-main.280>, <https://aclanthology.org/2022.emnlp-main.280>
47. Zang, X., Liu, L., Wang, M., Song, Y., Zhang, H., Chen, J.: PhotoChat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 6142–6152. Association for Computational Linguistics

- tics, Online (Aug 2021). <https://doi.org/10.18653/v1/2021.acl-long.479>, <https://aclanthology.org/2021.acl-long.479>
- 48. Zhang, Y., Sun, S., Galley, M., Chen, Y.C., Brockett, C., Gao, X., Gao, J., Liu, J., Dolan, B.: Dialogpt: Large-scale generative pre-training for conversational response generation. In: ACL, system demonstration (2020)
  - 49. Zhao, J., Zhang, T., Hu, J., Liu, Y., Jin, Q., Wang, X., Li, H.: M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5699–5710. Association for Computational Linguistics, Dublin, Ireland (May 2022). <https://doi.org/10.18653/v1/2022.acl-long.391>, <https://aclanthology.org/2022.acl-long.391>
  - 50. Zheng, K., He, X., Wang, X.E.: Minigpt-5: Interleaved vision-and-language generation via generative vokens. arXiv preprint arXiv:2310.02239 (2023)
  - 51. Zheng, Y., Chen, G., Liu, X., Sun, J.: MMChat: Multi-modal chat dataset on social media. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis, S. (eds.) Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 5778–5786. European Language Resources Association, Marseille, France (Jun 2022), <https://aclanthology.org/2022.lrec-1.621>

## A Limitations

Our framework relies on customized text-to-image models to ensure image consistency in multimodal dialogues. While these models generally offer better consistency than standard text-to-image models without conditioning, they are not infallible and may sometimes fail to accurately capture the conditioned input image. This represents a current limitation of our work. However, with the rapid advancements in customized text-to-image generation, we expect these shortcomings to decrease over time.

## B Broader Impact

It is crucial to emphasize that the main contribution of our work is not the customized text-to-image model itself but the overall framework that facilitates its effective use in multimodal dialogue scenarios. By focusing on enhancing image consistency, our framework opens up new avenues for more coherent and engaging multimodal interactions. This underscores the potential of our approach in revolutionizing how conversational agents handle multimodal inputs and responses, paving the way for more sophisticated and human-like dialogue systems.

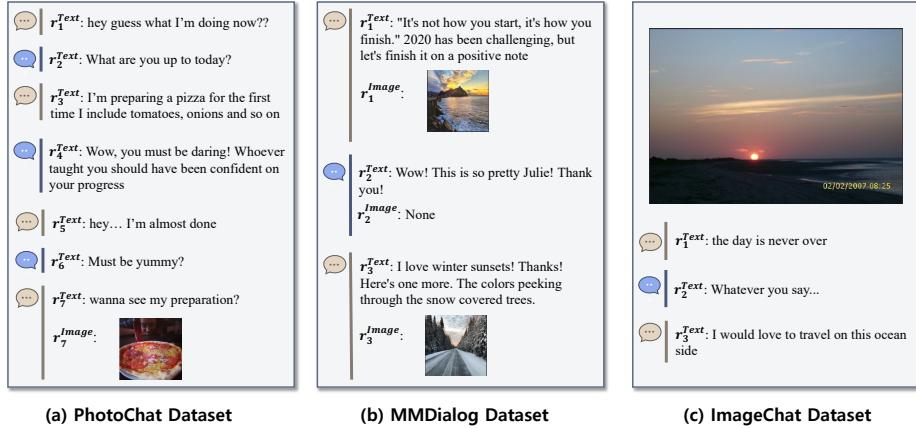
## C Benchmark Datasets

### C.1 Categorization of Existing Multimodal Dialogue Datasets.

As stated in Section 2.1, Multimodal dialogue datasets generally fall into three categories: question and answering (Q&A), in-scene, and conversation-based. In Table 5, we summarize the datasets for each category.

**Table 5:** Summary of Multimodal Dialogue Datasets. The type can generally be classified into three categories: question and answering (Q&A), the conversation taking place in a scene from a video (in-scene), and natural multimodal conversation (conversation-based). The modalities can contain audio (a), video (v), image (i), or text (t).

Dataset	Dialogue Type	Modalities	Dialogue Source	Turns	Language	Public
VisDial [8]	Q&A	i,t	crowd-sourcing	2.47M	English	o
AVSD [1]	Q&A	a,v,t	crowd-sourcing	236K	English	o
OpenViDial [31]	in-scene	i,t	movies&TVs	1.1M	English	o
OpenViDial 2.0 [44]	in-scene	i,t	movies&TVs	5.6M	English	o
YTD-18M [12]	in-scene	a,v,t	movies&TVs	5.6M	English	o
ImageChat [42]	conversation-based	i,t	crowd-sourcing	401K	English	o
PhotoChat [47]	conversation-based	i,t	crowd-sourcing	156K	English	o
MMDD [20]	conversation-based	i,t	text datasets	346K	English	o
DialogCC [21]	conversation-based	i,t	text datasets	929K	English	x
MMDialog [9]	conversation-based	i,t	social media	4.92M	English	o
MMChat [51]	conversation-based	i,t	social media	314K	Chinese	o
TikTalk [27]	conversation-based	a,v,t	social media	827K	Chinese	o



**Fig. 10: Example of Benchmark Datasets used in our paper.**

## C.2 PhotoChat Dataset

PhotoChat [47] features dialogues collected from social media, where a single image is shared in one of the conversation turns, which mirrors everyday human interaction. An example of PhotoChat dialogue is shown in Figure 10-(a).

## C.3 MMDialog Dataset

The limited scale and domain diversity of the PhotoChat dataset restricts its applicability. Overcoming these limitations, MMDialog [9] features over a million diverse dialogues from social media, where multiple images are shared across numerous conversation turns, providing a more realistic representation of open-domain multimodal conversations. An example of MMDialog dialogue is shown in Figure 10-(b).

## C.4 ImageChat Dataset

To evaluate the image-grounding advantage of our BI-MDRG to the previous system, we use the ImageChat Dataset [42]. This dataset has three turns of conversation about a given image. An example of ImageChat Dialogue is shown in Figure 10-(c).

## C.5 Multimodal Dialogue Image Consistency (MDIC) Dataset

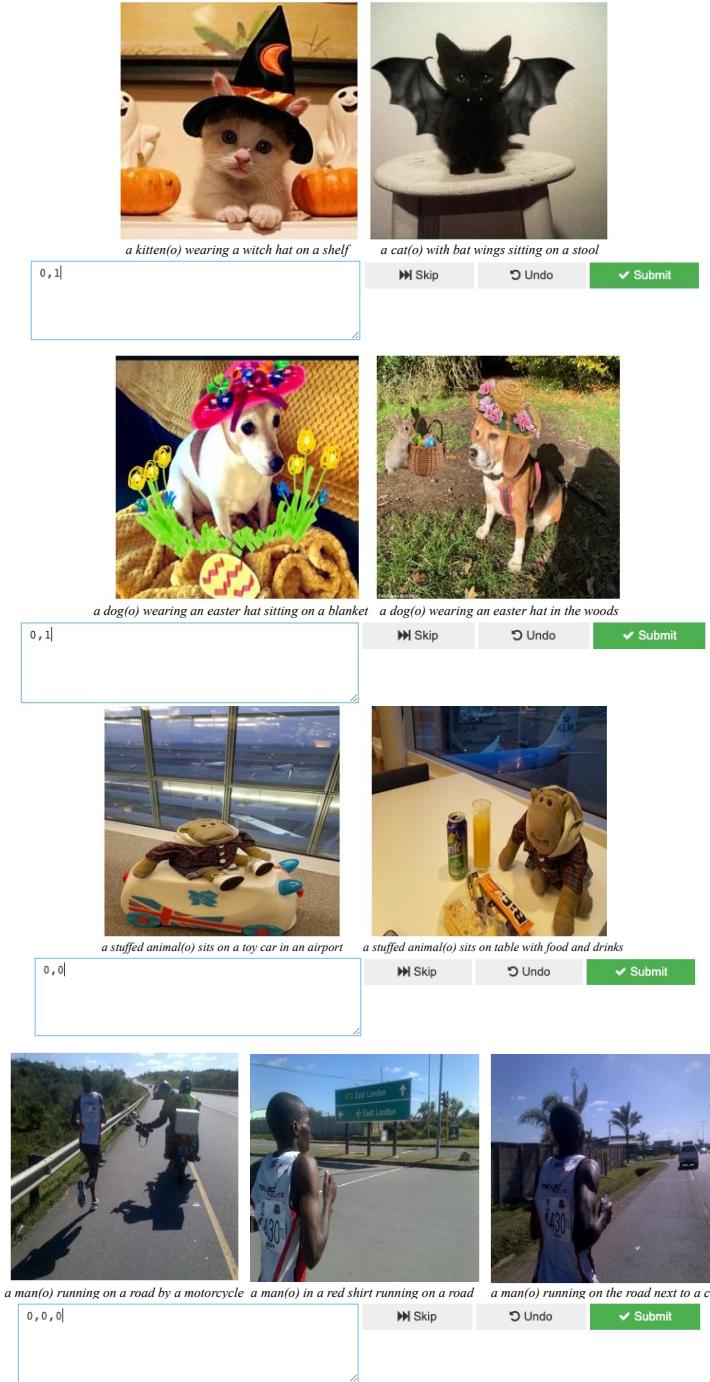
The challenge of ensuring consistent image generation in multimodal dialogue systems is amplified by the absence of datasets annotated for entity consistency across conversational images. We developed the Citation Module for our BI-MDRG system to address this gap. This module is designed to pseudo-label the



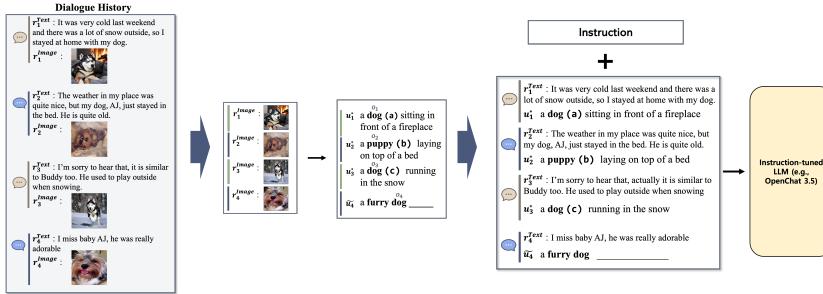
**Fig. 11:** Illustration of the labeling interface used for creating the Multimodal Dialogue Image Consistency (MDIC) dataset. The interface presents all images associated with a specific dialogue from the MMDialog test set. Labelers are tasked with assigning citation tags to the primary objects in these images, identified as *(o)*. The assignment is based on visual similarity and the identity of objects across different images.

recurring visual entities throughout a dialogue, allowing us to train our model to generate textual image descriptions during inference with citations that reflect the objects needing consistency. However, a benchmark dataset with explicit image consistency annotation is essential to validate the Citation Module and our BI-MDRG, which was trained with the pseudo-labels created from the Citation Module. To this end, we created the Multimodal Dialogue Image Consistency (MDIC) dataset. This dataset comprises a collection of dialogues annotated to identify the recurring visual entities across the conversation.

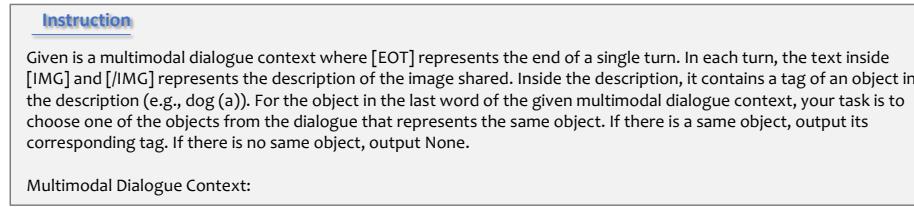
**Labeling Process** MDIC benchmark dataset was created using a labeling process applied to the images from the MMDialog test set. Figure 11 illustrates the labeling interface used. For each dialogue’s images, its corresponding textual image descriptions were obtained using BLIP2-flan-t5-xl [24] and pre-processed using spaCy [13] to identify the primary objects in the sentence. Five annotators examined these images and descriptions and assigned citation tags to the primary objects based on visual similarity and the identity of the objects across different images (examples of annotations are shown in Figure 12). For instance, if a dialogue contained two images with the same object, the labeler would input ‘0,0’; if the two images contained different objects, the labeler would input ‘0,1’. The final dataset selections were based on a consensus approach, retaining only those test sets where all five annotators unanimously agreed.



**Fig. 12:** Examples of labeled annotations of the MDIC dataset. The labeler inputs comma-separated numbers that represent the citation of the primary object in the textual image description based on the object's similarity.



**Fig. 13: Illustration of the LLMCite Baseline.** This approach employs an instruction-tuned large language model for assigning citation tags, treating citation tag prediction as a multiclass classification task. Specifically, it involves selecting the object from the dialogue history that the current object identically matches.



**Fig. 14: Instruction given to the LLM for the LLMCite baseline.**

## D Details on LLMCite

In Sections 4.4 and 5, we employ a baseline citation approach, LLMCite, illustrated in Figure 13, which leverages an instruction-tuned large language model (LLM) to assign citation tags (specifically, we use *OpenChat 3.5* (7B)<sup>4</sup>). From the MDIC dataset, we frame citation tag prediction as a multiclass classification task. Given a dialogue history  $D = \{(r_i^{\text{Text}}, r_i^{\text{Image}})\}_{i=1}^t$ , we first convert images into textual descriptions to form  $\{r_1^{\text{Text}}, u_1, \dots, r_t^{\text{Text}}, u_t\}$ . For the last turn  $t$ , we preprocess  $u_t$  to include only up to the principal object  $o_t$ , denoted as  $\tilde{u}_t$ . For preceding turns  $u_{1:t-1}$ , we append classification tags  $c_{1:t-1}^*$  (sequentially labeled as (a), (b), (c), ...) to principal objects  $o_{1:t-1}$ , resulting in augmented descriptions  $u_{1:t-1}^*$ . This modified sequence  $\{r_1^{\text{Text}}, u_1^*, \dots, r_t^{\text{Text}}, \tilde{u}_t\}$  is then provided to the LLM with instructions, as illustrated in Figure 14, to choose the most appropriate  $c_{1:t-1}^*$  matching  $o_t$  within the dialogue context.

<sup>4</sup> [https://huggingface.co/openchat/openchat\\_3.5](https://huggingface.co/openchat/openchat_3.5)

## E Additional Examples

In Section 5, we demonstrated that merely increasing the model size does not enhance image consistency. This limitation arises because the framework relies on text as an intermediary step for generating image responses, leading to an inherent loss of image information. ChatGPT also operates within this framework, utilizing text as an intermediary due to the challenges and infeasibility of implementing an end-to-end model, a point underscored in Section 5. Consequently, our proposed framework, specifically designed to maintain image consistency, becomes critical. Figure 15 illustrates that ChatGPT also struggles to maintain image consistency, reinforcing the need for our targeted framework.

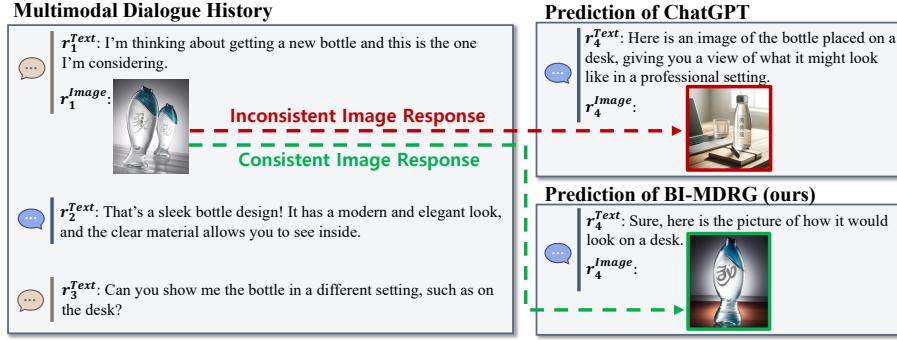


Fig. 15: Image Response of ChatGPT and BI-MDRG (ours).

In Figures 16 and 17, we present further examples of BI-MDRG predictions, showcasing both image-grounded textual responses and the model’s ability to maintain consistency in image responses.

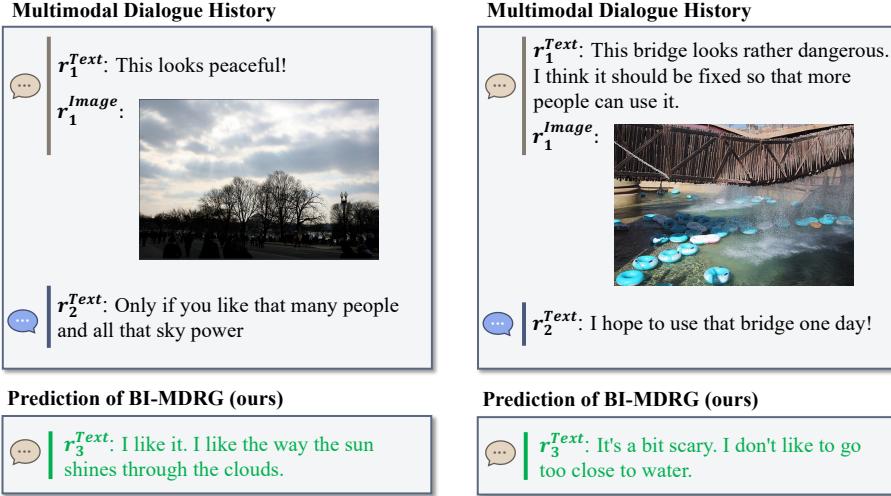


Fig. 16: Examples of Image-Grounded Text Response of BI-MDRG (ours).

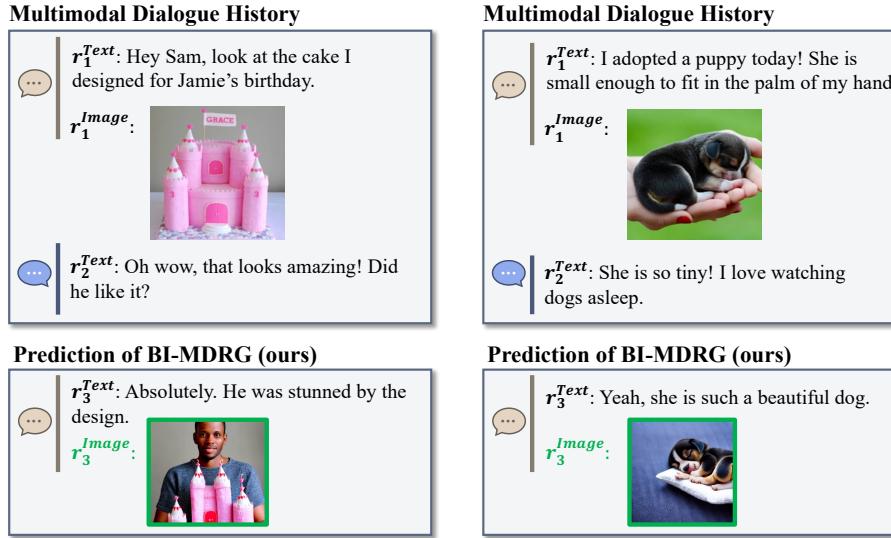


Fig. 17: Examples of Consistent Image Response of BI-MDRG (ours).