

## 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 [48]	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 [52]	conversation-based	i,t	social media	314K	Chinese	o
TikTalk [27]	conversation-based	a,v,t	social media	827K	Chinese	o

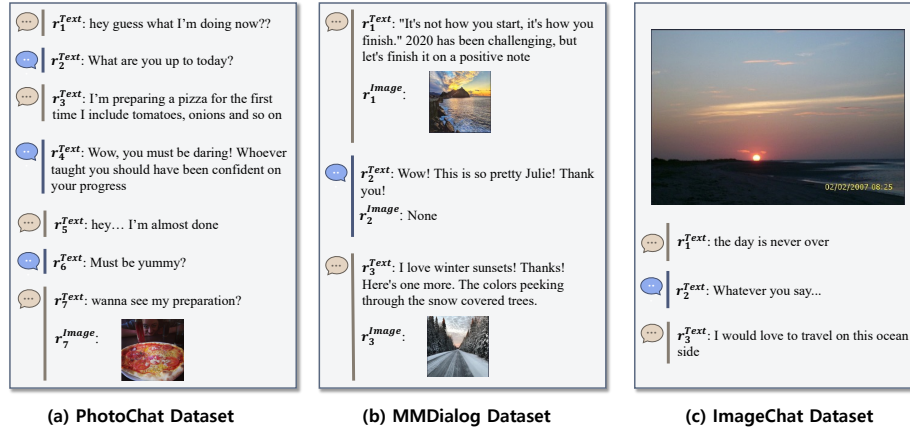


Fig. 10: Example of Benchmark Datasets used in our paper.

## C.2 PhotoChat Dataset

PhotoChat [48] 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

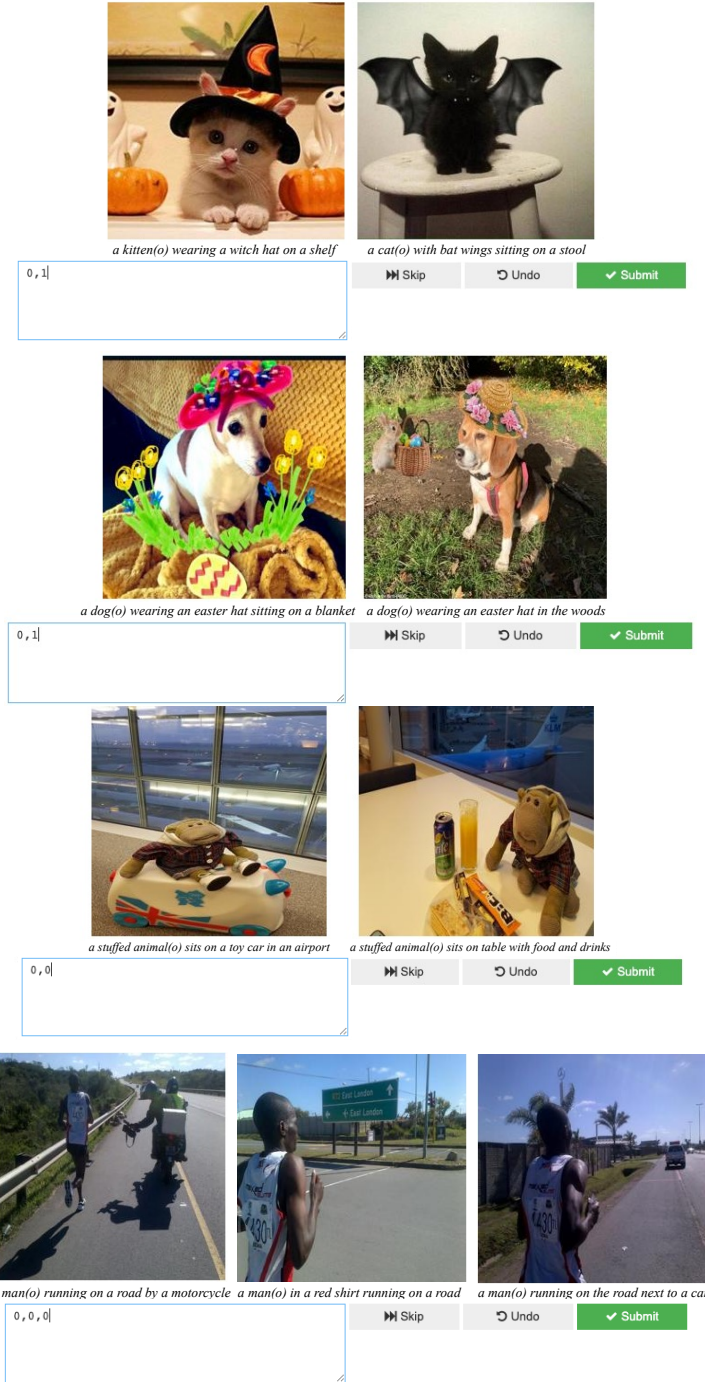
The challenge of ensuring consistent image generation in multimodal dialogue systems is amplified by the absence of datasets annotated for entity consistency across conversational images. We developed the Citation Module for our BI-MDRG system to address this gap. This module is designed to pseudo-label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



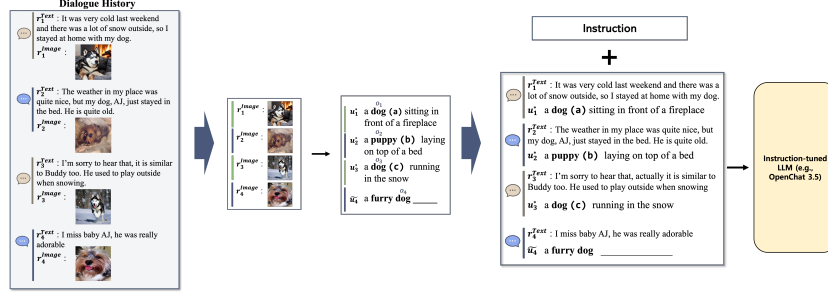
**Fig. 11:** Illustration of the labeling interface used for creating the Multimodal Dialogue Image Consistency (MDIC) dataset. The interface presents all images associated with a specific dialogue from the MMDialog test set. Labelers are tasked with assigning citation tags to the primary objects in 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.

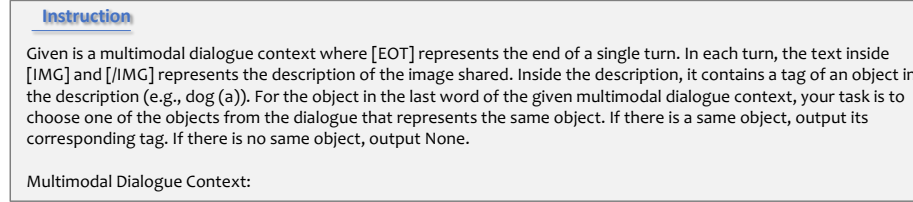
**Labeling Process** MDIC benchmark dataset was created using a labeling process applied to the images from the MMDialog test set. Figure 11 illustrates the labeling interface used. For each dialogue’s images, its corresponding textual image descriptions were obtained using BLIP2-flan-t5-xl [24] and 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.

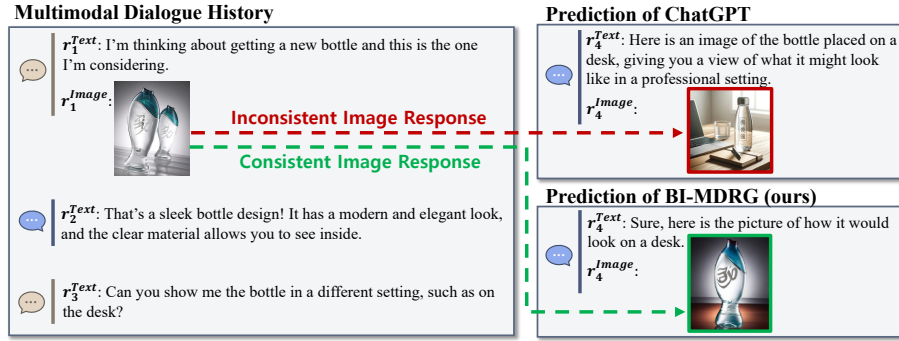



Fig. 15: Image Response of ChatGPT and BI-MDRG (ours).

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.

**Multimodal Dialogue History**

...  $r_1^{Text}$ : This looks peaceful!

$r_1^{Image}$ : 


...  $r_2^{Text}$ : Only if you like that many people and all that sky power

**Prediction of BI-MDRG (ours)**

...  $r_3^{Text}$ : I like it. I like the way the sun shines through the clouds.

**Multimodal Dialogue History**

...  $r_1^{Text}$ : This bridge looks rather dangerous. I think it should be fixed so that more people can use it.

$r_1^{Image}$ : 

...  $r_2^{Text}$ : I hope to use that bridge one day!

**Prediction of BI-MDRG (ours)**

...  $r_3^{Text}$ : It's a bit scary. I don't like to go too close to water.

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

**Multimodal Dialogue History**


...  $r_1^{Text}$ : Hey Sam, look at the cake I designed for Jamie's birthday.

$r_1^{Image}$ : 

...  $r_2^{Text}$ : Oh wow, that looks amazing! Did he like it?

**Prediction of BI-MDRG (ours)**

...  $r_3^{Text}$ : Absolutely. He was stunned by the design.

$r_3^{Image}$ : 

**Multimodal Dialogue History**

...  $r_1^{Text}$ : I adopted a puppy today! She is small enough to fit in the palm of my hand

$r_1^{Image}$ : 

...  $r_2^{Text}$ : She is so tiny! I love watching dogs asleep.

**Prediction of BI-MDRG (ours)**

...  $r_3^{Text}$ : Yeah, she is such a beautiful dog.


$r_3^{Image}$ : 

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