MAGID: An Automated Pipeline for Generating Synthetic Multi-modal Datasets

Hossein Aboutalebi[†,⚑] **Hwanjun Song**[‡,♠] **Yusheng Xie**[a] **Arshit Gupta**[a] **Justin Sun**[a] **Hang Su**[a] **Igor Shalyminov**[a] **Nikolaos Pappas**[a] **Siffi Singh**[a] **Saab Mansour**[a]

[⚑] Cheriton School of Computer Science, University of Waterloo

[♠]Korea Advanced Institute of Science and Technology

[a] AWS AI Labs

haboutal@uwaterloo.ca

Abstract

Development of multimodal interactive systems is hindered by the lack of rich, multimodal (text, images) conversational data, which is needed in large quantities for LLMs. Previous approaches augment textual dialogues with retrieved images, posing privacy, diversity, and quality constraints. In this work, we introduce **M**ultimodal **A**ugmented **G**enerative **I**mages **D**ialogues (MAGID), a framework to augment text-only dialogues with diverse and high-quality images [1]. Subsequently, a diffusion model is applied to craft corresponding images, ensuring alignment with the identified text. Finally, MAGID incorporates an innovative feedback loop between an image description generation module (textual LLM) and image quality modules (addressing aesthetics, image-text matching, and safety), that work in tandem to generate high-quality and multi-modal dialogues. We compare MAGID to other SOTA baselines on three dialogue datasets, using automated and human evaluation. Our results show that MAGID is comparable to or better than baselines, with significant improvements in human evaluation, especially against retrieval baselines where the image database is small.

1 Introduction

In recent years, advancements in large language models (LLMs) have expanded possibilities and research directions in AI, with studies highlighting their extensive capabilities in handling dialogue datasets (Liu et al., 2023c; Penedo et al., 2023). Specifically, there is a growing interest in their application to multi-modal dialogue datasets, given that *sharing images* is an integral aspect of human-human conversations (Alayrac et al., 2022; OpenAI, 2023; Liu et al., 2023a).

Several multi-modal dialogue datasets like MMDialog (Feng et al., 2022), DialogCC (Lee et al., 2022), and PhotoChat (Zang et al., 2021) have been introduced for training multi-modal LLMs. These datasets either use a retrieval-based approach, pulling images from set image banks, such as MS-COCO (Lin et al., 2014), or restrict the dialogue to only one image per conversation, even if they involve real human-human chats. Moreover, when leveraging real-world datasets from platforms like social media, issues related to privacy concerns and image quality become significant challenges for training.

As a result, these methods limit the diversity of images since the small image database cannot adequately capture the wide range of real human-human conversations (Lee et al., 2021, 2022). Additionally, they face challenges stemming from low-quality images containing harmful and private content (Feng et al., 2022) and shortage of accessible data (Lee et al., 2022), particularly when utilizing real human-human conversations from social media sources.

To address these challenges, we propose **MAGID**, a *generative*-based multi-modal dialogue creation framework. As illustrated in Figure 1, MAGID aims at converting existing text-only data into context-enriched multi-modal data by addressing the two research challenges: (i) how to find the most suitable utterances that can be enhanced by adding images and (ii) how to generate realistic and diverse images that do not have harmful and private contents.

[1]https://e77p.short.gy/MAGID

[†]Lead author. Work done as an intern at AWS AI Labs.

[‡]Corresponding author; Work done while at AWS AI Labs.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5150–5167

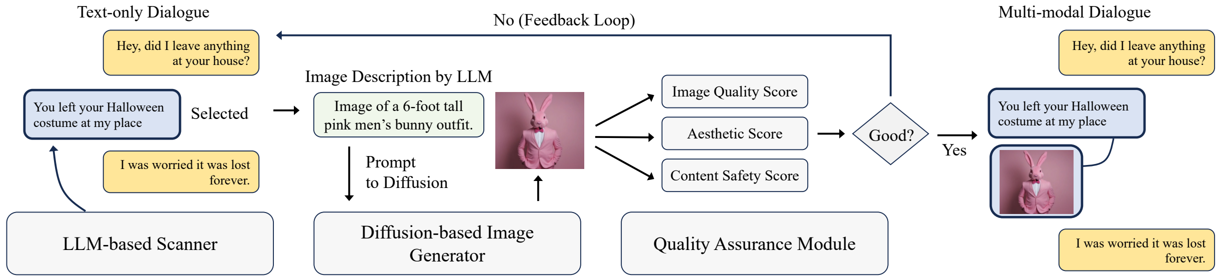


Figure 1: Overview of the MAGID framework. MAGID consists of three components: (1) LLM-based scanner to identify suitable utterances to augment with images, (2) diffusion-based image generator to create realistic images, and (3) quality assurance module to enhance the image quality, aesthetic and safety scores. The text-only dialogue is automatically converted to multi-modal dialogue using MAGID.

In the former case, we introduce an *LLM-based scanner* designed to pinpoint utterances requiring images and subsequently generate corresponding image descriptions, leveraging chain-of-thought prompting. In the latter case, we employ a *diffusion-based image generator*, adept at crafting images with notable diversity, drawing upon the generated image descriptions as its input. Additionally, a *quality assurance* module is incorporated into our framework to ensure both the congruence and the quality of the produced images, thereby preserving coherence and fidelity within the multi-modal dialogue. Should the generated image not satisfy the criteria of this module, MAGID initiates a feedback loop, revisiting the processes of prompt and image generation.

Distinct from numerous previous endeavors that have depended on image-retrieval techniques for curating multi-modal datasets (Lee et al., 2021, 2022)—a method that might result in restricted image diversity and potential mismatch with the dialogue existing utterances—we employ the generative model Stable Diffusion XL (Podell et al., 2023). By training on billions of images (Schuhmann et al., 2022), this approach guarantees an output that is both rich and varied. Such outputs align well with the conversational context provided by the LLM feedback, thereby elevating the quality and diversity of our multi-modal dataset.

Our framework aligns with prior studies using text-only datasets (Lee et al., 2021, 2022), but it addresses the limitations associated with their retrieval-based strategies by employing a generative-based data creation method. Unlike Liu et al. (2023a); Lee et al. (2021), we do not restrict the inclusion of only one image per dialogue. Consequently, MAGID generates synthetic yet more realistic multi-modal dialogue datasets thus mitigating data accessibility issues and facilitating the development of advanced multi-modal models.

To summarize, our main contributions are:

- We present MAGID, a generative-based multi-modal dialogue data creation framework that addresses the limitation of retrieval-based approaches.
- We conduct experiments using various prompt engineering strategies to optimize interactions between the LLM-based scanner and the diffusion-based image generator.
- We propose a novel quality assurance design to control the performance of generative models effectively.
- We provide a medium-sized dataset as a proof of concept to showcase the effectiveness of MAGID pipeline (section 5).
- We conduct extensive human evaluations on the dataset and test multiple LLM models to ensure robustness and reliability.

2 Related Works

2.1 Generative Models

Recent advances in Generative AI has started new trends in expanding capabilities of existing deep learning models. In NLP, works like (Radford et al., 2019; Ouyang et al., 2022) have shown importance of training data to build better LLM models. In this regard, recent LLM models like Falcon-40b-Instruct (Penedo et al., 2023), Koala 13b (Geng et al., 2023), LLaMA 13b (Touvron et al., 2023),

Zero shot prompt

You are an AI assistant that helps augment textual dialogues with engaging images. As input, you will receive a conversation between people which is represented as a sequence of utterances. As output, you will generate a description of images that can support the utterances in the conversation.
The format of the input is 'Utterance i: ...' where 'i' denotes the order of the Utterance in the conversation. Given this query, you output in the format of
<result>Utterance i: image_description</result> <reason>explanation_of_choice </reason>
where 'i' is the Utterance in the conversation and 'image_description' is the short text description of an image that can be followed by that Utterance that can make the conversation more engaging. You should only identify the most appropriate utterances in the conversation.
The text inside <reason>explanation_of_choice</reason> is the explanation of why you picked the utterance with the image description.

Figure 2: The zero-shot prompt of the scanner module (Section 3.1) which selects turns in the dialogue to augment with images and generates descriptions of those images. Additional few-shot and chain-of-thought prompts are provided in the supplementary materials (section A).

OpenLLaMA (Touvron et al., 2023), and Vicuna 13b (Chiang et al., 2023) use better curated training datasets to achieve higher performances. In this regard, paper like Christiano et al. (2017) has shown the dramatic impact of using higher quality data (from human feedback) in faster training. Yet, using human feedback and crowd-sourcing is not always cheap. To address this, emerging works like (Veselovsky et al., 2023; Kamalloo et al., 2023) suggests that LLM has the capabilities of performing the task of human generated dataset. In addition, diffusion models in computer vision have shown promising results in generating images indistinguishable from real ones (Podell et al., 2023; Ho et al., 2020). Finally, recent works focus on building multi-modal LLM models including GPT-4 (OpenAI, 2023), LLaVA (Liu et al., 2023b), AnyMAL(Moon et al., 2023) which supports any modality. Specifically, LLaVA accepts multi-modal input, combining image and text embeddings to generate text-only output.

2.2 Multi-modal Dataset Creation

There are also works which focus on generating multi-modality datasets. In particular, MMDD (Lee et al., 2021) and DialogCC (Lee et al., 2022) use image-retrieval approaches to augment text-only datasets to multi-modal datasets. PhotoChat (Zang et al., 2021) hires workers to discuss a particular image to build the dataset. MMDialog (Feng et al., 2022) collect multi-modal conversations from internet to build the dataset which can potentially pose privacy concern to use as training set. There are also works (Wang et al., 2023; Corona et al., 2021, 2020; Ciliberto et al., 2021; Abdrakhmanova et al., 2021) which focuses modality beyond text and image including video and voice. For example, Corona et al. (2021) provide a dataset that contains videos for activity detection. IntenVid (Wang et al., 2023) is another example that contains video in addition to text.

3 MAGID Pipeline

In transitioning from text-only to multi-modal dialogue, there exist two core challenges. The first is the identification of the most suitable utterances within the dialogue that can be enhanced by images. The second is the creation of corresponding, accurate images that align with the selected utterances. In this regard, we need to ensure a harmonious and coherent match between the image and the text, achieving acceptable image-text alignment.

We have addressed these challenges through the implementation of the following three key modules in Figure 1, namely LLM-based scanner, diffusion-based image generator, and quality assurance module, which are detailed in the subsequent sections.

3.1 MAGID Scanner

The primary objective of this module is to identify suitable utterances that can be visually represented by an image. Achieving best performance requires precise control over the behavior of the LLM model. We use prompt engineering and special formatting to control the output of LLM.

We experimented with three prompt engineering strategies to fine-tune the system prompts of the LLM:

- **Zero-shot prompting:** The LLM is provided with only the format of the input and the expected output, along with a general problem description. Figure 2 shows an example of the zero-shot prompt.

Scanner (GPT-4 Answer):

<reason>

Utterance 0 is asking for recipes and offering a gift card as a prize. An image of the gift card can be used to attract more participants.

Utterance 1 does not require an image as it is a response to the previous utterance.

Utterance 2 is asking someone to look at something, so an image of what they want the other person to look at would be appropriate.

</reason>

<result>

Utterance: 0: An image of a gift card worth INR 500

Utterance: 2: An image of the thing the speaker wants the other person to look at

</result>

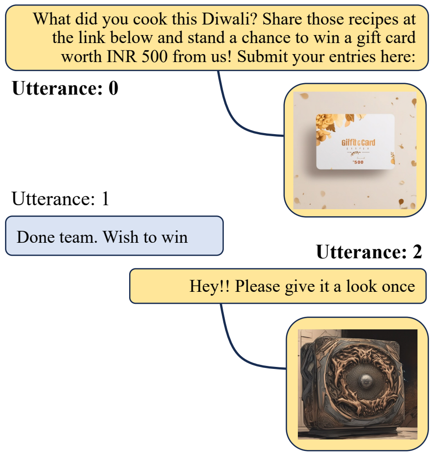


Figure 3: MAGID's chain of thought prompting facilitates debugging and identification of corner cases, utilizing the SDXL 1.0 diffusion model and GPT-4 (OpenAI, 2023). The depicted conversation is sourced from a real human-human interaction in the MMDialog dataset (Feng et al., 2022).

- **Few-shot example prompting:** Besides the information provided in zero-shot prompting, LLM is also supplied with several input–output exemplars to demonstrate the anticipated response from the LLM model (Brown et al., 2020). We have included this type of prompt in supplementary materials (section A).

- **Chain of Thought prompting:** As per (Wei et al., 2022), this prompting strategy involves imparting a series of intermediate reasoning steps for each example, facilitating the LLM model's capacity for more advanced reasoning. Please refer to supplementary materials for example of this prompt (section A).

In section 4.3.1, we evaluated these prompting strategies. Based on the findings, we selected Chain of Thought prompting as the optimal choice for our MAGID framework.

3.2 Controlling LLM Output Format

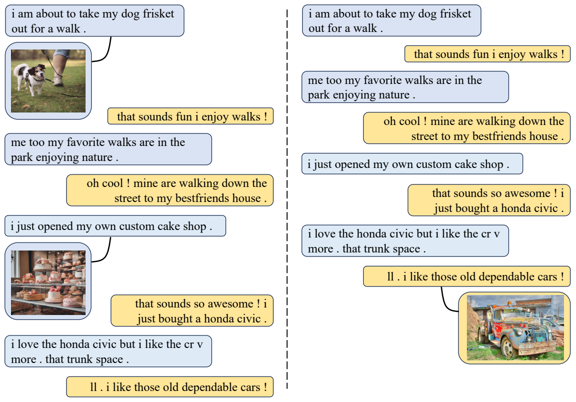
We introduce a method that seeks to streamline the structuring of LLMs outputs by employing HTML-like tags, aiming to facilitate easier parsing and to shed light on the decision-making process. The utilization of $< result >$ and $< reason >$ tags is intended to envelope answers and rationales respectively, potentially making post-processing more straightforward and offering a degree of transparency into the model's reasoning, which may be beneficial for debugging purposes.

Figure 3 demonstrates the impact of using the proposed HTML formatting inside chain of thought prompt, revealing how meticulous analysis of responses identifies corner cases and ensures contextual congruency in produced images. Whereas the first image aligns with preceding text, the second lacks context. The $< reason >$ tag discloses that phrases like "give it a look" influenced image generation. To enhance contextual relevance and model reliability, the system prompt has been refined to instruct the LLM to only generate images when paired with a detailed description, thereby avoiding contextual discrepancies.

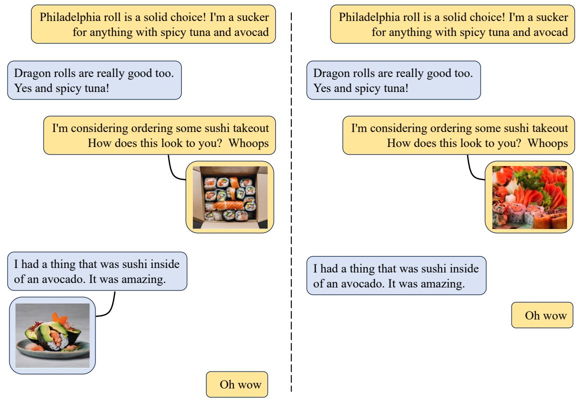
3.3 MAGID Image Generator

As illustrated in Figure 1, the LLM model's image prompts are used by the diffusion model to generate corresponding images. In this regard, given the success of diffusion models in superior image generation (Rombach et al., 2022; Ho et al., 2020), were chosen over GANs (Goodfellow et al., 2014). Models tested included SDXl 1.0, SDXL 0.9, and Stable Diffusion versions from Stability AI (Podell et al., 2023), with a detailed comparison in supplementary materials (section C).

Ultimately, SDXl 1.0 was chosen for its state-of-the-art capabilities, bolstering the quality and reliability of the generated images of the MAGID dataset. Nevertheless, future model developments can be incorporated to refine our MAGID dataset generation.



(a) MAGID (left) vs. MMDD (right).



(b) MAGID (left) vs. PhotoChat (right).

Figure 4: Qualitative comparison of MAGID with an image retrieval-based synthetic MMDD and a real human image-based PhotoChat datasets.

3.4 MAGID Quality Assurance

The Quality Assurance (QA) module is essential for improving the MAGID pipeline's efficiency. It assures the generated images satisfy user-set standards in three domains: **Image-Text Matching**, **Image Quality**, and **Image Safety**.

1- Image-text Matching: We use the CLIP score (Radford et al., 2021) to validate the match between the image and the LLM model's utterance. A low CLIP score triggers image regeneration, with the count determined as a hyperparameter. In this work, we set the regeneration count to two.

2- Image Quality: Images are rated based on an aesthetic score from (Schuhmann et al., 2022; Schuhmann, 2023), which uses CLIP embedding followed by an MLP. This model identifies artifacts in the diffusion model outputs. A threshold of 0.51 efficiently detects most artifacts, prompting image regeneration for scores below this.

3- Image Safety: Image safety, particularly against NSFW content, is crucial. While many models assess this, few unsafe images were found in our dataset, indicating our process's reliability.

This robust QA ensures that MAGID can output relevant, high-quality, and safe images.

3.4.1 Feedback Loop

Should the diffusion model produce an image that does not meet the quality assurance module's stipulations, the issues might stem from the LLM model's prompt. Faulty prompts can yield low image-text matches or unsafe images. To mitigate this, our design, showcased in Figure 1, includes a feedback loop, instructing the LLM model to generate a better image description given regenerated images with previous image description continuously fall short of quality assurance standards.

Figure 4 displays a comparison of MAGID samples with two other datasets, MMDD (Lee et al., 2021) and PhotoChat (Zang et al., 2021). A qualitative analysis shows MAGID yields quality comparable to real datasets, such as PhotoChat, and surpasses synthetic datasets like MMDD in generating high-quality multi-modal dataset. More examples are included in supplementary (section H).

4 Evaluation

We scrutinize the efficacy and applicability of the multi-modal dataset generated by MAGID. Here are three pivotal questions we addressed in evaluation:

1. How does MAGID quantitatively compare against real multi-modal datasets? ▷ Section 4.1
2. Can MAGID create a multi-modal dataset with human-eye perceptible quality like a real one? ▷ Section 4.2
3. What is the impact of scanner prompt tuning and the quality assurance module on MAGID? ▷ Section 4.3

The first and third question delves into a quantitative analysis, probing the accuracy and quality of the data generated by MAGID. Moreover, the second question is crucial, as a failure of MAGID to meet human evaluation standards would result in a low-quality training dataset that is unable to get positive human-centric assessments.

In addition, in supplementary (section E), we have studied training multimodal model

Table 1: Scanner module performance as measured by turn selection for image augmentation (accuracy, precision, recall, F1) and the resulting images from the generated descriptions (CLIP, MM-relevance, aesthetic) on the MMDialog dataset as ground-truth. The quality assurance module is **enabled**. We compare various LLMs powering the scanner module using chain of thought prompting.

Model	Accuracy	Precision	Recall	F1 score	CLIP score	MM-Relevance	Aesthetic	#images
GPT 4	**67.24%**	**70.49%**	**46.87%**	**0.56**	**0.27**	**294.52**	0.57	1359
GPT 3.5	63.54%	69.43%	33.97%	0.46	0.26	293.51	**0.58**	1001
Falcon-40b-Ins.	58.93%	61.26%	24.13%	0.35	0.25	254.50	**0.58**	794
Koala 13b	56.28%	62.33%	6.91%	0.12	0.25	243.31	0.57	223
Llama 13b	57.10%	60.00%	13.64%	0.22	0.25	247.99	0.57	460
OpenLLaMA	57.94%	64.36%	12.69%	0.21	0.25	250.96	**0.58**	390
Vicuna 13b	58.77%	66.60%	14.38%	0.24	0.26	255.18	0.57	506
MMDialogue[2]	N/A	N/A	N/A	N/A	0.262	N/A	0.47	2717

with MAGID and compared it with using real images for training.

4.1 Quantitative Evaluation

Setup. Addressing the first question, a multi-dimensional evaluation assessed the image quality and accuracy of MAGID in selecting right utterances. To fairly compare MAGID's general-use applicability, we only utilized prompt engineering to guide the LLM model to select the right utterances. In this regard, as a ground truth, we selected human-human interaction datasets MMDialog and PhotoChat, and removed images from their test sets and employed MAGID to transform the text-only data into a multi-modal dataset.

For the LLM-based model, we adopted a range of models, including GPT-4 (OpenAI, 2023), GPT-3.5 (OpenAI, 2023), Falcon-40b-Instruct (Penedo et al., 2023), Koala 13b (Geng et al., 2023), LLaMA 13b (Touvron et al., 2023), OpenLLaMA (Touvron et al., 2023), and Vicuna 13b (Chiang et al., 2023). For image generation, SDXL 1.0 was consistently utilized across all models. We present the results of the MMDialog dataset here, and the PhotoChat results are included in supplementary (section B). In these experiments, we have set the threshold for the CLIP model at 0.21 and the aesthetic score threshold of 0.51. We used grid search to find these hyperparameters. More details on computational cost is provided in supplementary (section F).

Result. Table 1 presents the performance of various LLM models on the MMDialog dataset. The table quantifies MAGID's response generation using different LLM models in comparison to the MMDialog dataset. The first column lists the LLM models used, while the subsequent four columns measure accuracy, precision, recall, and F1 score in choosing the correct utterance to be augmented with an image. The CLIP score gauges image-text matching, and the MM-Relevance, as introduced in (Feng et al., 2022), denotes the similarity between responses. In our context, it determines the resemblance of the produced image to the MMDialog's original image. The next column, the aesthetic score, indicates the image quality as discussed in (Schuhmann, 2023). Last row presents the ground truth dataset, highlighting the CLIP score, image count, and aesthetic quality of its images.

From the table, it is evident that GPT-4 and GPT-3.5 outperforms other models across all metrics. Notably, the CLIP and aesthetic scores of MAGID using GPT-4 and GPT-3.5 surpass even the ground truth values. In the next section, we also examine image-text matching and image quality in our human evaluation for MAGI against other datasets to test if it is aligned with our quantitative findings.

4.2 Human Evaluation

Setup. We conducted a human evaluation using a website with questionnaire. Participants viewed two dialogues: one with an image from MAGID and another from datasets MMDD (Lee et al., 2021), PhotoChat (Zang et al., 2021), or MMDialog (Feng et al., 2022). MAGID used GPT-4 as its Language Model and SDXL 1.0 for image generation. From the mentioned datasets, we selected 20 dialogues each, totaling 60 dialogues, and replaced their images with MAGID's. During evaluation, participants compared MAGID's multi-modal dialogues with the originals, without informa-

Table 2: Human Evaluation results of MAGID created datasets versus a retrieval-based synthetic dataset, MMDD, and two real datasets, MMDialouge and PhotoChat, where the mean shows the percentage of time the dialogues in one dataset were preferred among participants. (Q1: more realistic dialogue? Q2: images in which dialogue provide more knowledge?, Q3: better text-image matched?, Q4: better context-image matched?, Q5: more engaging?, Q6: higher image quality?)

	(a) MAGID vs. MMDD			(b) MAGID vs. MMDialogue			(c) MAGID vs. PhotoChat		
#	Mean MAGID	Mean MMDD	Gwet's AC1	Mean MAGID	Mean MMDial.	Gwet's AC1	Mean MAGID	Mean Photo.	Gwet's AC1
Q1	**96.29**%	3.71%	0.74	48.17%	**51.83**%	0.63	**58.11**%	41.89%	0.47
Q2	**96.29**%	3.71%	0.89	49.33%	**50.67**%	0.65	**68.24**%	31.76%	0.71
Q3	**89.11**%	10.89%	0.75	**52.72**%	47.28%	0.54	**64.90**%	35.10%	0.53
Q4	**91.11**%	8.89%	0.83	46.31%	**53.69**%	0.65	**61.98**%	38.02%	0.54
Q5	**95.57**%	4.43%	0.89	**51.94**%	48.06%	0.63	**64.02**%	35.98%	0.61
Q6	**80.92**%	19.08%	0.65	**63.90**%	36.10%	0.55	**69.99**%	30.01%	0.64

Table 3: Utterance selection accuracy using three different prompts on MMDialogue (ground-truth), where ZS, FS, and CoT stand for zero-shot, few-shot, and chain of thought respectively.

Prompt	Accuracy	Precision	Recall	F1 score
ZS	65.53%	73.12%	36.16%	0.48
FS	63.89%	69.67%	34.45%	0.46
CoT	**68.51**%	**73.37**%	**47.32**%	**0.57**

tion about the dialogue origins.

For each dialogue pair (one from MAGID and one from the benchmark datasets), participants responded to the following prompts:

Q1: Which dialogue appears more realistic?

Q2: Which dialogue's images convey greater knowledge?

Q3: In which dialogue is there better match between images and the immediately preceding text?

Q4: In which dialogue do the images more closely match with the overall conversation context?

Q5: Which dialogue is more engaging?

Q6: Which dialogue features higher quality images?

Respondents selected from binary choices (Dialogue A or Dialogue B) for each prompt.

For this evaluation, 15 human annotators provided their answers. Schema of the website interface are available in the Supplementary materials (section D).

Result. Table 2 displays MAGID's results against MMDD, MMDialog, and PhotoChat datasets. The 'Mean MAGID' column shows the percentage of annotators favoring MAGID, while 'Mean Other' indicates those preferring the alternative dataset. Gwet's AC1 measure, found in the last column, was used to assess inter-annotator reliability. It offers stability over Cohen's Kappa (Wongpakaran et al., 2013) and is more resilient to outliers (For more explanation, please refer to Supplementary Materials section G.).

From Table 2(a), it's evident that annotators favored MAGID over the synthetically generated MMDD dataset across all question categories. Moreover, the high Gwet's AC1 value indicates a strong consensus among annotators in choosing MAGID over MMDD. In contrast, when examining Table 2(b), annotators exhibited a slight preference for the authentic MMDialog dataset in terms of realism. Notably, the Gwet's AC1 value is considerably lower here than in the MMDD results, suggesting a reduced consensus among annotators. Nevertheless, MAGID outperformed MMDialog in terms of image quality and image-text matching. **Such findings affirm our quantitative evaluations and showcase the potential of generative AI in producing superior data sources for training.** As for the PhotoChat dataset (Table 2(c)), while it is constructed from authentic human interactions, human participants were told to mock real conversation. Interestingly, our annotators slightly leaned towards MAGID over PhotoChat. This outcome suggests MAGID's promising capability to serve as an alternative to Mechanical Turk in the development of multi-modal datasets.

4.3 Ablation Study of MAGID

We conducted ablation studies on (1) using different prompts for utterance identification

Table 4: Ablation results of the MAGID framework with and without the quality assurance (QA) module. Results on turn selection and image quality performance across four LLMs on MMDialog (ground-truth) are shown. The first four rows are the results with the QA module, while the last four are the results without. The system prompt is chain of thought.

Model	Accuracy	Precision	Recall	F1 score	CLIP score	MM-Relevance	Aesthetic	#images
GPT 4	**67.24%**	**70.49%**	**46.87%**	**0.56**	**0.27**	**294.52**	0.57	1359
GPT 3.5	63.54%	69.43%	33.97%	0.46	0.26	293.51	**0.58**	1001
Falcon-40b-Ins.	58.93%	61.26%	24.13%	0.35	0.25	254.50	0.58	794
OpenLLaMA	57.94%	64.36%	12.69%	0.21	0.25	250.96	**0.58**	390
GPT 4	67.86%	69.70%	**50.64%**	**0.59**	**0.27**	**282.25**	0.55	1485
GPT 3.5	**68.51%**	**73.37%**	47.32%	0.57	0.26	278.16	0.55	1109
Falcon-40b-Ins.	56.77%	53.58%	28.80%	0.37	0.23	224.59	0.55	1075
OpenLLaMA	58.92%	62.50%	21.51%	0.32	0.21	213.56	**0.56**	696

and (2) investigating the impact of our quality assurance (QA) module.

4.3.1 Prompts for Scanner

Table 3 displays the outcomes of three prompt strategies, namely Zero-shot (ZS) prompting, Few-shot prompting (FS), and Chain of Thought (CoT) prompting, as applied to the GPT-3.5 model for MAGID. These results are reported for the MMDialog dataset, with quality assurance deactivated, to solely measure the accuracy of the LLM model. Notably, the Chain of Thought strategy outperforms the other two across all evaluated metrics.

4.3.2 Impact of QA Module

Table 4 showcases the performance of four LLM models in MAGID, contrasting when the QA module is either enabled or disabled. A perusal of Table 4 reveals a decline in the aesthetic score, MM-Relevance, and CLIP score across all models upon the deactivation of QA. Moreover, a noticeable decrement in the precision of most models is observable, validating that the QA module bolsters MAGID by enhancing precision in pinpointing the optimal utterance for image generation. In contrast, disabling QA leads to an elevation in recall, attributable to MAGID selecting a more extensive array of utterances for image generation, thereby reducing the ratio of false negatives. Future research could explore the development of a refined QA module capable of elevating the recall rate for the entire pipeline.

5 MAGID Dataset

As a proof of concept, and consistent with studies like (Lee et al., 2021), we employed text-only datasets such as DailyDialog (Li et al., 2017), Persona-Chat (Zhang et al., 2018), and PhotoChat (Zang et al., 2021) (by

Table 5: Statistics of the MAGID dataset.

Category	Train	Test
Total dialogues	53071	5208
Avg length of dialogues	8.53	11.37
Avg length of sentences	10.75	9.99
Total images	75654	8938

replacing its images with MAGID) to generate a multi-modal dataset [3] of 58,279 dialogues. Based on the results of our experiments, we used GPT-3.5 to transform 52,527 input dialogues and GPT-4 to augment the rest. Table 5 shows the statistics of the generated dataset with MAGID.

6 Conclusion

We presented a generative, fully automated pipeline, MADIG, designed to transform text-only datasets into multi-modal variants, harnessing the power of LLMs through prompt engineering. This solution addresses limitations faced by preceding methods, notably in terms of data privacy, accessibility, constrained image distribution, and occurrences of unsuitable or non-consensual content. Crucially, our pipeline permits the substitution of real, potentially privacy-compromising images with synthetic counterparts. We thoroughly evaluated MAGID using human assessment, quantitative analyses with various LLMs, and an in-depth ablation study. The promising results highlight generative AI's capability to stand as an alternative to traditional data generation methods, like mechanical turk.

Looking ahead, our dataset paves the way for developing large multi-modal language models that can engage with users via both text and visuals.

[3] The link to dataset: https://e77p.short.gy/MAGID

Limitations

This paper predominantly concentrates on augmenting the privacy, diversity, and quality of multi-modal dataset generation by employing LLM and diffusion models. Although utilizing generative diffusion models can mitigate issues related to privacy breaches—given these models are also trained on extensive volumes of web images—they are susceptible to copyright infringement (Aboutalebi et al., 2023). Addressing this issue exceeds the scope of this paper and presents a compelling avenue for future work.

Moreover, the current work exclusively emphasizes image and text modalities. Extending considerations to additional modalities—such as video sharing, voice sharing, and more—is recommended for subsequent research endeavors. In addition, fine-tunning of large language model to generate image is left to future works.

Improving generated image consistency in the dialogue is another important aspect that can further improve the quality of the generated multi-modal dataset by MAGID. Employing more recent diffusion models such as DALL-E 3 (Betker et al., 2023) can address this problem as they can make more consistent image generation. In this regard, in the section J of Supplementary materials, we have included further examples that shows the limitations of the proposed MAGID pipeline.

In conclusion, the enhancement of our quality assurance module is pivotal for developing more realistic multi-modal datasets from text-only inputs. In this regard, works like (Tian et al., 2023) already showed that using synthesized images is effective. This work prioritizes aspects like aesthetic score, clip score, and safety. Future research can explore additional elements to further refine and add realism to the transformation into multi-modal outputs.

Broader Impact

This research introduces a novel approach aimed at enhancing privacy during the training phase of multi-modal language models through the strategic use of Generative AI. By employing this methodology, our framework effectively filters out images that could potentially contain sensitive or personal information, thereby ensuring a more ethical and responsible model training process. Central to our approach is the application of generative models for image synthesis and the integration of a Large Language Model (LLM) module designed to uphold the highest standards of impartiality and ethical compliance. This involves a stringent adherence to guidelines that prevent the generation of harmful or biased data.

Our initiative is grounded in the belief that by leveraging a synthetic data framework, we can significantly alleviate concerns related to privacy violations often associated with the training of large-scale language models. This strategy not only contributes to the creation of safer AI systems but also paves the way for the development of AI technologies that are fundamentally more responsible to the ethical implications of their deployment.

References

Madina Abdrakhmanova, Askat Kuzdeuov, Sheikh Jarju, Yerbolat Khassanov, Michael Lewis, and Huseyin Atakan Varol. 2021. Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams. *Sensors*, 21(10):3465.

Hossein Aboutalebi, Daniel Mao, Carol Xu, and Alexander Wong. 2023. Deepfakeart challenge: A benchmark dataset for generative ai art forgery and data poisoning detection. *arXiv preprint arXiv:2306.01272*.

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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion

Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Mathias Ciliberto, Vitor Fortes Rey, Alberto Calatroni, Paul Lukowicz, and Daniel Roggen. 2021. Opportunity++: A multimodal dataset for video-and wearable, object and ambient sensors-based human activity recognition. *Frontiers in Computer Science*, 3:792065.

Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. 2020. Meva: A large-scale multiview. *Multimodal Video Dataset for Activity Detection*.

Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. 2021. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1060–1068.

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

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

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. *Blog post, April*, 1.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

Gant Laborde. Deep nn for nsfw detection.

Min Young Lee. 2023. Building multimodal ai chatbots. *arXiv preprint arXiv:2305.03512*.

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

Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, and Ho-Jin Choi. 2022. Dialogcc: Large-scale multi-modal dialogue dataset. *arXiv preprint arXiv:2212.04119*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. *arXiv preprint arXiv:2202.13587*.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023c. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

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

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Christoph Schuhmann. 2023. improved-aesthetic-predictor. https:https://github.com/christophschuhmann/improved-aesthetic-predictor. GitHub repository.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. 2023. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *arXiv preprint arXiv:2306.00984*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13:1–7.

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

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Supplementary

A COT & FS Prompts

In the paper, we referenced the **Few Shot** and **Chain of Thought** prompts, which can be found in Figures 5 and 6, respectively. When generating multi-modal versions from each text-only input dataset, it became evident that distinct prompting is necessary for the chain of thoughts due to variations in the format of the input text.

B PhotoChat results

As mentioned in section 4.1, here we have included the results of different LLM on PhotoChat dataset. Table 6 shows the results. Overall, GPT 3.5 shows better performance compared with other LLM models. As it can be seen, the precision is significantly lower compared with the results reported on MMDialogue dataset (Table 1) and that is because this dataset is limited to only one image per dialogue while our pipeline does not have such restriction.

Table 6: Different LLM model testing on PhotoChat (ground-truth). Quality Assurance module is enabled. The system prompt is chain of thoughts.

Model	Accuracy	Precision	Recall	F1 score	CLIP score	MM-Relevance	#images	Aesthetic
GPT 3.5	86.11%	**28.62%**	25.91%	0.27	0.25	313.64	87	0.57
Falcon-40b-Ins.	88.10%	28.04%	11.83%	0.17	0.24	303.68	403	**0.58**
Koala 13b	89.61%	30.43%	2.94%	0.05	0.24	283.44	92	0.61
Llama 13b	87.32%	20.79%	9.54%	0.13	0.23	244.36	433	0.59
OpenLLaMA	88.75%	27.31%	8.03%	0.12	0.23	270.36	696	0.59
Vicuna 13b	88.40%	25.48%	8.35%	0.13	0.24	244.97	602	0.55
PhotoChat	N/A	N/A	N/A	N/A	0.213	N/A	961	0.49

Few-shot example prompt

```
- query: >
    Utterance: 0: So yeah, it was a mostly dismal year. But what's the best news you've
    read/saw/heard in 2016? (Anything from the personal to world affairs.)
    Utterance: 1: grew 14 pumpkins on the formidable  strength of my chickens We are
    all proud! Here's one
    Utterance: 2: Very impressive!
  answer: >
    <result> Utterance: 1: 14 pumpkins</result>
- query: >
    Utterance: 0: Working from home with a tie today! Plenty of Zoom in my life today!
    Utterance: 1: I keep a polo handy that I throw on and off for zoom calls.
    Way to be extra fancy
  answer: >
    <result>Utterance: 0: Working from home with a tie</result>
```

Figure 5: The few-shot example prompt not only provides the format for both input and expected output along with a problem description but also includes multiple exemplars to elucidate the desired response from the LLM. Here only exemplars are included.

Chain of Thoughts prompt

```
- query: >
    Utterance: 0: So yeah, it was a mostly dismal year. But what's the best news you've
    read/saw/heard in 2016? (Anything from the personal to world affairs.)
    Utterance: 1: grew 14 pumpkins on the formidable  strength of my chickens We are
    all proud! Here's one
    Utterance: 2: Very impressive!
  answer: >
  <reason> Utterance 0 is just a descrption of last year without any information that
    can be translated with image. We add photographic style as it is a personal sharing.
    Utterance 1 on the other hand is talking about growing 14 pumpkins. This can be
    represented with image.</reason>
    <result> Utterance: 1: 14 pumpkins</result>
- query: >
    Utterance: 0: My attire for the SA Hip Hop Awards
    Utterance: 1: Are you a supporter of Kaizer Chiefs?...lol. Gorgeous!!
  answer: >
    <reason>In Utterance 0 contains the sentence "My outfit for
    the SA hip hop awards" which shows
    the person is willing to share her outfit</reason>
    <result>Utterance: 0: My outfit for the SA hip hop awards </result>
```

Figure 6: The chain of thoughts prompt, building upon the system prompt provided in the few-shot example prompt, also incorporates the detailed reasoning on utterance selection.

C Image Generator Ablation Study

Table 7 shows the performance of different diffusion models (Podell et al., 2023; Rombach et al., 2022). The results are taken from MMDialog dataset and the quality assurance module is disabled to report the results without filtering unwanted ones. It is clear that SDXL 1.0 and SDXL 0.9 has very similar performance and higher aethetic score compared with Stable Diffusion 2.0. All models have similar CLIP score which is predictable as they are given the same prompt for image generation.

D Human evaluation

To collect answers from annotators, we created a website with a schema shown in Figure 7. For each question, annotators were given two screenshots of the same dialogue, one generated by MAGID and the other from a source dataset (PhotoChat, MMDialog, or MMDD). At the start of the annotation session, annotators were instructed to ignore the conversation text and focus only on the images and image-text matching. Fifteen annotators completed the task, each making 20 comparisons.

E Downstream Training

Here, we study how much MAGID can impact training a multi-modal model when changing the original image with synthetic one generated by MAGID. In addition, we also compare it with benchmark cases when no image is present in the training and with MMDD (Lee et al., 2021) approach to include image in the dialogue. In this regard, we used the same architecture suggested in (Lee, 2023) which is visionTextDualEncoder from Huggingface (Wolf et al., 2019) which projects the encoding of image with the the embedding of text to a shared common space. For encoding of image we used ViT (Dosovitskiy et al., 2020), and for processing the text we used pretrained DialoGPT (Zhang et al., 2019). While the input is multimodal, the output is text only. In this task, we omit the last text utterance and the model should predict it given the prior image and text.

Table 7: Testing different Stable diffusion models with MAGID pipeline

Model	Aesthetic Score	CLIP Score
SDXL 1.0	0.56	0.26
SDXL 0.9	**0.57**	0.26
Stable Diffusion 2.0	0.53	0.26

We fine-tuned the model on MMDialog dataset and the results are reported in Table 8. For this experiment, we used the learning rate of 0.00005 with Adam Optimizer. In Table 8, we show the results on the test set when training set images is coming from MMDialogue, MAGID, MMDD and the case where the images are omitted. For MMDD, we used the same code they used to inject image into text-only dialogue to make the comparison possible. For this expeiment, the training set consists of 5156 dialogues and the test set consists of 633 dialogues sampled from MMDialogue dataset.

Table 8: Downstream training. The model used is DialoGPT + ViT. BLUE score is in percentage.

Dataset	PPL	BLEU-1	distinct-1	distinct-2
MMDialogue	73.09	**8.3**	0.94	0.96
MAGID	**70.99**	7.9	0.94	**0.97**
MMDD	78.86	7.5	0.93	0.96
No image	78.88	7.9	0.92	0.95

As it can be seen, when we use the source image as training set (MMDialog), we achieve highest BLEU score (Papineni et al., 2002). The perplexity of the model using MAGID is lowest which shows the model is more confident in making the prediction. In addition, the distinct score (Liu et al., 2022) which shows the diversity of response is highest with MAGID which can be attributed to higher image-text match provided with MAGID images. It is important to note that since MMDialog dataset is a real dataset, the quality of images shared does not necessarily matches the text and this can make the model less confident and results in higher perplexity. On the other hand, the images generated by MAGID is more controlled.

For this experiment we used 4 NVIDIA RTX GPU each with 24 GiB memory and the training took for a full day.

F Experiment Computational Cost

For running MAGID pipeline, it can be run with one GPU with NVIDIA RTX with 24 GiB memory.

G Discussion on Inter-rater Reliability Measure Choice

In Section 4.2, we employed Gwet's AC1 for evaluating the consistency among reviewers, opting not to use Cohen's Kappa due to its susceptibility to outliers and potential for showing inconsistent results despite high average scores across all participants. As detailed in the study by Wongpakaran et al. (2013), Gwet's AC1 is recognized for its greater consistency in inter-rater reliability assessments when compared to Cohen's Kappa, alongside its enhanced resilience to outliers, providing a more reliable measure for our analysis (Wongpakaran et al., 2013). This approach ensures a more stable and accurate assessment of reviewer consistency, mitigating the impact of anomalies on the reliability scores.

H Further examples of MAGID

Figures 8, 9, and 10 provide more examples on comparing MAGID with MMDialog, PhotoChat, and MMD.

I Experiment Setting

For determining the threshold for image-text matching and aesthetic score, we employed cross-validation on the validation set. In this regard, the threshold for CLIP score was set for 0.21 and the threshold for the aesthetic score was set for 0.51. Based on our observations, we established a protocol where a generated image could fail up to two times before being discarded and triggering the feedback loop. This approach ensured a balance between generating high-quality images and maintaining efficient processing. In all our experiments, we used SDXL 1.0 model for image generation. Finally, For image safety and NSFW detection, the library from (Laborde) can be used.

J Limitations

In Figures 11, 12, and 13, we showcase the most common scenarios were MAGID can fail to generate the image which properly supports the preceding utterance. Specifically, figure 11 shows a common example, where the generated image usually fails to put the proper text sign in the generated image. In Figures 12 and 13 showcase the examples where the generated image does not follow the correct description in terms of number object should exist in the image. We believe using more advanced diffusion models like DALL-E 3 should mitigate this problem.

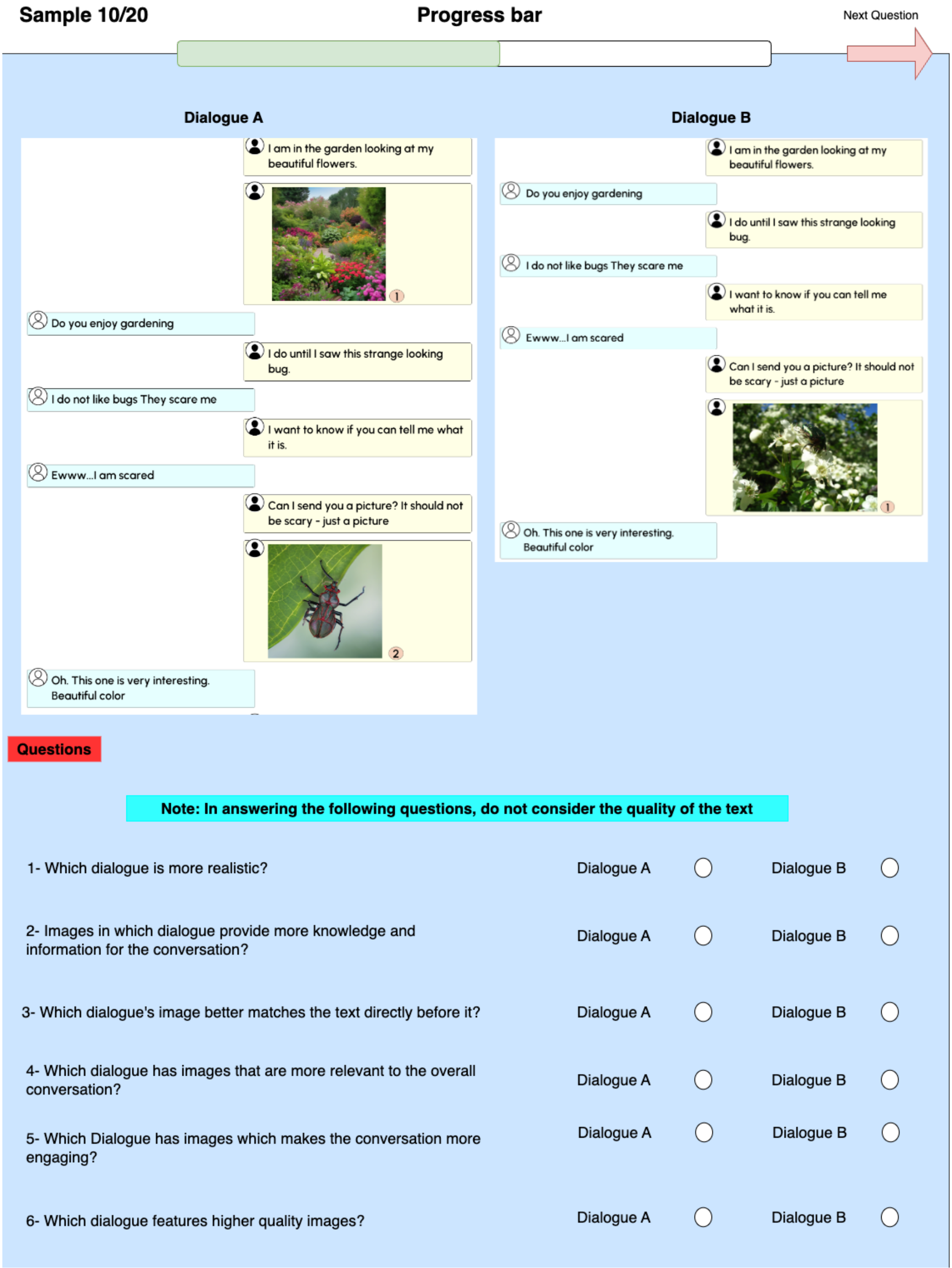


Figure 7: Schema of the website used to perform human evaluation.

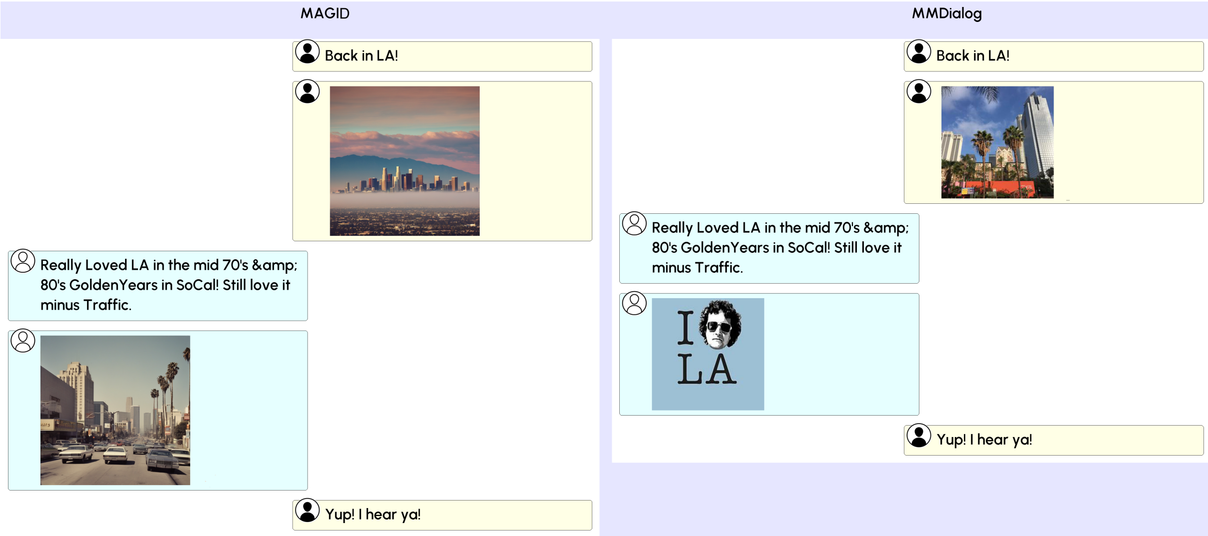


Figure 8: MAGID (left) versus MMDialog (right)

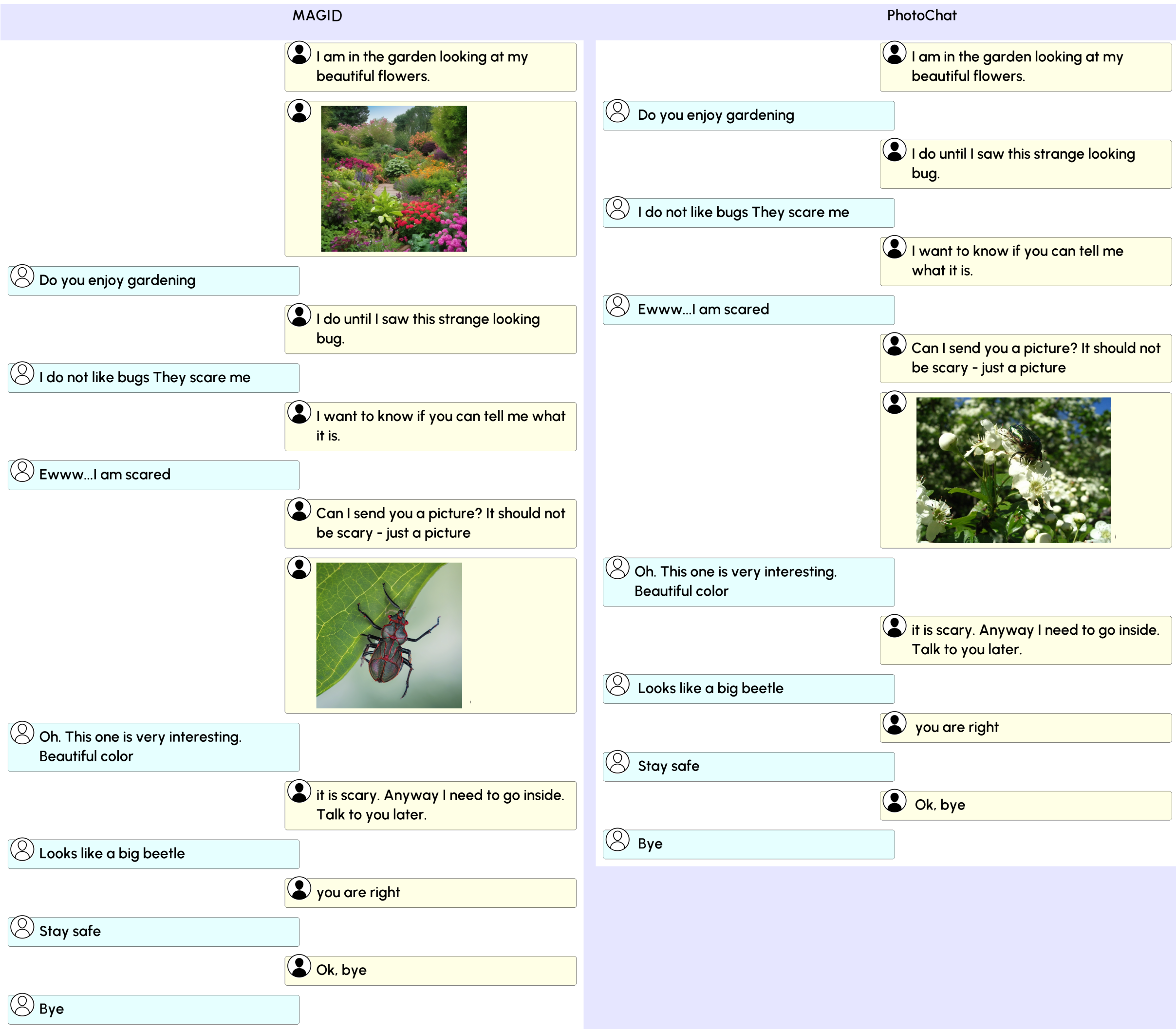


Figure 9: MAGID (left) versus PhotoChat (right)

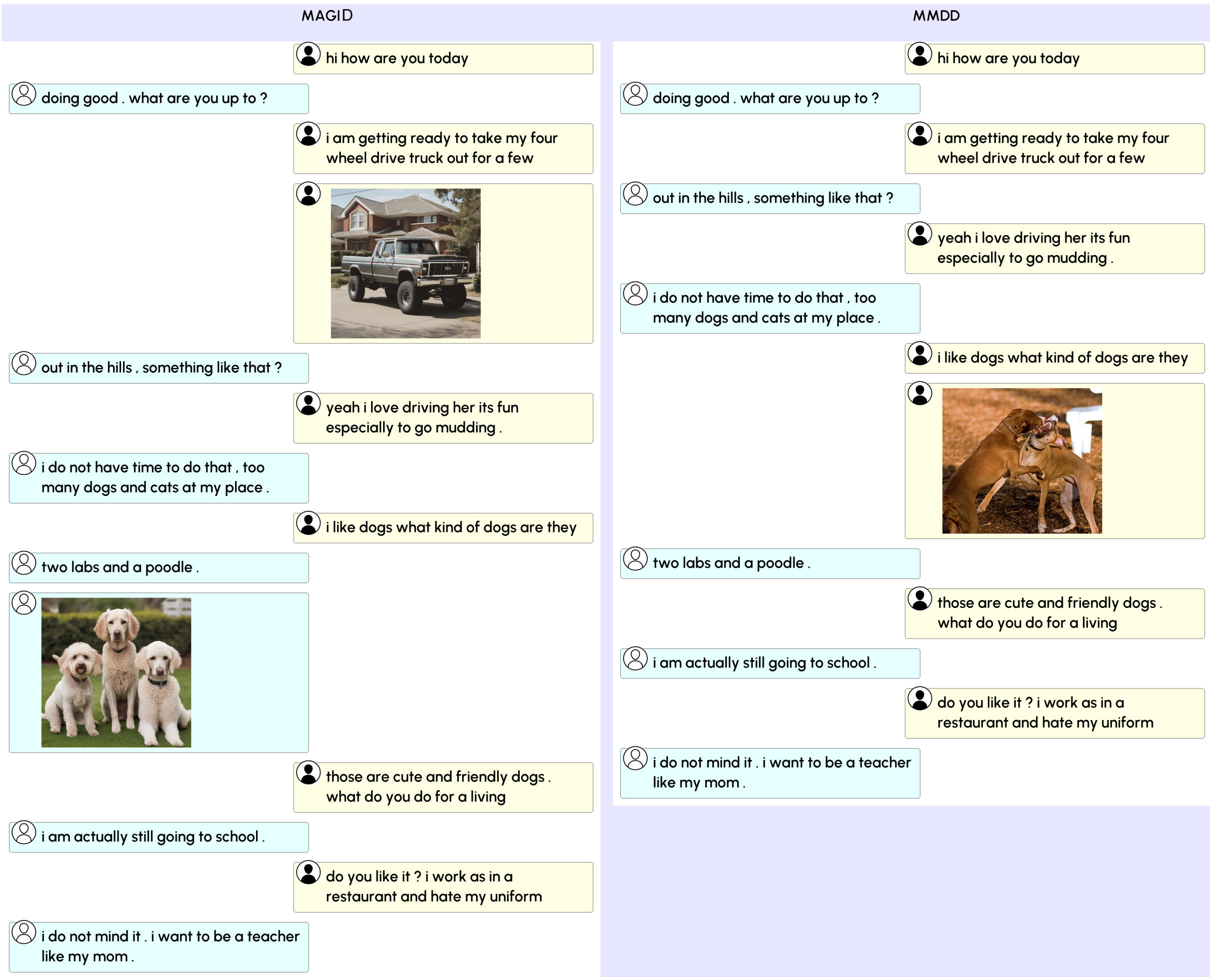


Figure 10: MAGID (left) versus MMDD (right)

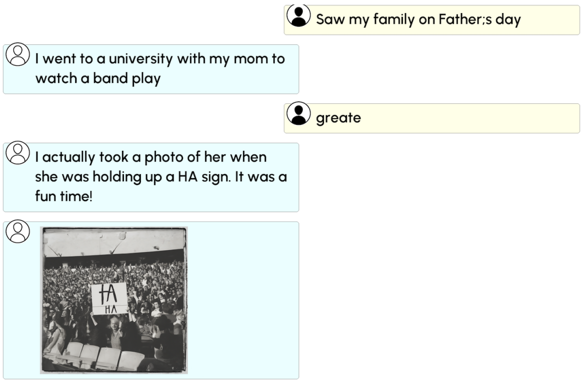


Figure 11: Generated image by MAGID fails to properly show the sign HA

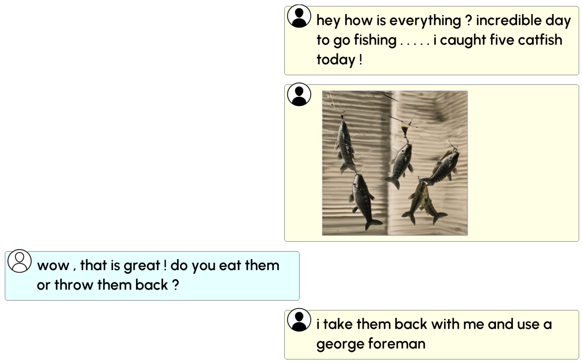


Figure 13: Generated image by MAGID fails to properly shows 5 fishes instead of 6

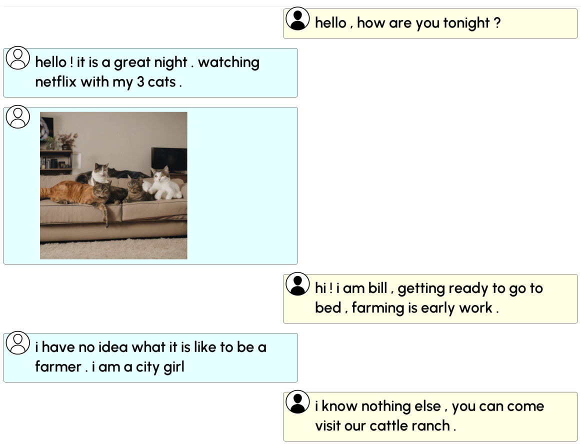


Figure 12: Generated image by MAGID fails to properly shows 3 cats instead of 4