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

Hossein Aboutalebi†, Hwanjun Song‡, Yusheng Xie§, Arshit Gupta§, Justin Sun§, Hang Su§, Igor Shalyminov§, Nikolaos Pappas§, Siffi Singh§, Saab Mansour§

†Cheriton School of Computer Science, University of Waterloo
‡Korea Advanced Institute of Science and Technology
§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 Multimodal Augmented Generative Images Dialogues (MAGID), a framework to augment text-only dialogues with diverse and high-quality images. 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 multimodal 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 multimodal 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.
You left your Halloween costume at my place
Hey, did I leave anything at your house?
I was worried it was lost forever.

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 (Schulmann 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 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.
What did you cook this Diwali? Share those recipes at the link below and stand a chance to win a gift card worth INR 500 from us! Submit your entries here:

Utterance 0
An image of a gift card worth INR 500

Utterance 1
Done team. Wish to win

Utterance 2
Hey!! Please give it a look once

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.
3.4 MAGID Quality Assurance
The Quality Assurance (QA) module is essential for improving the MAGID pipeline’s efficacy. 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?  
2. Can MAGID create a multi-modal dataset with human-eye perceptible quality like a real one?  
3. What is the impact of scanner prompt tuning and the quality assurance module on MAGID?  

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
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 hyper-parameters. 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?)

<table>
<thead>
<tr>
<th>#</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>MAGID</td>
<td>Mean</td>
<td>MMDD</td>
<td>Gwet’s AC1</td>
<td>Mean</td>
<td>MAGID</td>
</tr>
<tr>
<td>1</td>
<td>96.29%</td>
<td>48.17%</td>
<td>51.83%</td>
<td>0.63</td>
<td>58.11%</td>
<td>41.89%</td>
</tr>
<tr>
<td>2</td>
<td>96.29%</td>
<td>49.33%</td>
<td>50.67%</td>
<td>0.65</td>
<td>50.67%</td>
<td>31.76%</td>
</tr>
<tr>
<td>3</td>
<td>89.11%</td>
<td>52.72%</td>
<td>47.28%</td>
<td>0.54</td>
<td>58.24%</td>
<td>35.10%</td>
</tr>
<tr>
<td>4</td>
<td>91.11%</td>
<td>46.31%</td>
<td>53.69%</td>
<td>0.65</td>
<td>61.98%</td>
<td>38.02%</td>
</tr>
<tr>
<td>5</td>
<td>95.57%</td>
<td>51.94%</td>
<td>48.06%</td>
<td>0.63</td>
<td>64.90%</td>
<td>35.98%</td>
</tr>
<tr>
<td>6</td>
<td>80.92%</td>
<td>63.90%</td>
<td>36.10%</td>
<td>0.55</td>
<td>69.99%</td>
<td>30.01%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZS</td>
<td>65.53%</td>
<td>73.12%</td>
<td>36.16%</td>
<td>0.48</td>
</tr>
<tr>
<td>FS</td>
<td>63.89%</td>
<td>69.67%</td>
<td>34.45%</td>
<td>0.46</td>
</tr>
<tr>
<td>CoT</td>
<td>68.51%</td>
<td>73.37%</td>
<td>47.32%</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Result. Table 2 displays MAGID’s results against MMDD, MMDialouge, 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 MMDialouge 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 MMDialouge 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>CLIP score</th>
<th>MM-Relevance</th>
<th>Aesthetic</th>
<th>#images</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT 4</td>
<td>67.24%</td>
<td>70.49%</td>
<td>46.87%</td>
<td>0.56</td>
<td>0.27</td>
<td>294.52</td>
<td>0.57</td>
<td>1359</td>
</tr>
<tr>
<td>GPT 3.5</td>
<td>63.54%</td>
<td>69.43%</td>
<td>33.97%</td>
<td>0.46</td>
<td>0.26</td>
<td>293.51</td>
<td>0.58</td>
<td>1001</td>
</tr>
<tr>
<td>Falcon-40b-Ins.</td>
<td>58.93%</td>
<td>61.26%</td>
<td>24.13%</td>
<td>0.35</td>
<td>0.25</td>
<td>254.50</td>
<td>0.58</td>
<td>794</td>
</tr>
<tr>
<td>OpenLLaMA</td>
<td>57.94%</td>
<td>64.36%</td>
<td>12.69%</td>
<td>0.21</td>
<td>0.25</td>
<td>250.96</td>
<td>0.58</td>
<td>390</td>
</tr>
</tbody>
</table>

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 MMDialoq 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 replacing its images with MAGID) to generate a multi-modal dataset 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.

<table>
<thead>
<tr>
<th>Category</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total dialogues</td>
<td>53071</td>
<td>5208</td>
</tr>
<tr>
<td>Avg length of dialogues</td>
<td>8.53</td>
<td>11.37</td>
</tr>
<tr>
<td>Avg length of sentences</td>
<td>10.75</td>
<td>9.99</td>
</tr>
<tr>
<td>Total images</td>
<td>75654</td>
<td>8938</td>
</tr>
</tbody>
</table>

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.
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-tuning 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


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


Gant Laborde. Deep nn for nsfw detection.


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


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 MMDi-}

5160
Table 6: Different LLM model testing on PhotoChat (ground-truth). Quality Assurance module is enabled. The system prompt is chain of thoughts.

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

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.

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 MMDial 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 aesthetic 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, MMDial, 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 MMDial (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.

We fine-tuned the model on MMDial 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 experiment, 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PPL</th>
<th>BLEU-1</th>
<th>distinct-1</th>
<th>distinct-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMDialogue</td>
<td>73.09</td>
<td>8.3</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td>MAGID</td>
<td>70.99</td>
<td>7.9</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>MMDD</td>
<td>78.86</td>
<td>7.5</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>No image</td>
<td>78.88</td>
<td>7.9</td>
<td>0.92</td>
<td>0.95</td>
</tr>
</tbody>
</table>

As it can be seen, when we use the source image as training set (MMDial), 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 MMDialogue 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
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 MMDialg, 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 where 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 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)
<table>
<thead>
<tr>
<th>MAGID</th>
<th>MMDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>In how are you today</td>
<td>In how are you today</td>
</tr>
<tr>
<td>doing good, what are you up to?</td>
<td>doing good, what are you up to?</td>
</tr>
<tr>
<td>I am getting ready to take my four wheel drive truck out for a few</td>
<td>I am getting ready to take my four wheel drive truck out for a few</td>
</tr>
<tr>
<td>out in the hills, something like that?</td>
<td>out in the hills, something like that?</td>
</tr>
<tr>
<td>I do not have time to do that, too many dogs and cats at my place</td>
<td>yeah I love driving her its fun especially to go mudding</td>
</tr>
<tr>
<td>I like dogs what kind of dogs are they</td>
<td>I like dogs what kind of dogs are they</td>
</tr>
<tr>
<td>two labs and a poodle</td>
<td>two labs and a poodle</td>
</tr>
<tr>
<td></td>
<td>those are cute and friendly dogs, what do you do for a living</td>
</tr>
<tr>
<td>I am actually still going to school</td>
<td>I am actually still going to school</td>
</tr>
<tr>
<td></td>
<td>do you like it? I work as in a restaurant and hate my uniform</td>
</tr>
<tr>
<td>do you like it? I want to be a teacher like my mom</td>
<td>do you like it? I want to be a teacher like my mom</td>
</tr>
</tbody>
</table>

Figure 10: MAGID (left) versus MMDD (right)
Figure 11: Generated image by MAGID fails to properly show the sign HA

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

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