Large Language Models can *Share* Images, Too!

Young-Jun Lee[1] **Dokyong Lee**[2] **Joo Won Sung**[2] **Jonghwan Hyeon**[1] **Ho-Jin Choi**[1]
[1] School of Computing, KAIST [2] KT Corporation
{yj2961, jonghwanhyeon, hojinc}@kaist.ac.kr {dokyong.lee, jwsung}@kt.com

Abstract

This paper explores the *image-sharing* capability of Large Language Models (LLMs), such as GPT-4 and LLaMA 2, in a zero-shot setting. To facilitate a comprehensive evaluation of LLMs, we introduce the PHOTOCHAT++ dataset, which includes enriched annotations (i.e., *intent*, *triggering sentence*, *image description*, and *salient information*). Furthermore, we present the gradient-free and extensible **D**ecide, Desc**ribe**, and **R**etrieve (DRIBER) framework. With extensive experiments, we unlock the *image-sharing* capability of DRIBER equipped with LLMs in zero-shot prompting, with ChatGPT achieving the best performance. Our findings also reveal the emergent *image-sharing* ability in LLMs under zero-shot conditions, validating the effectiveness of DRIBER. We use this framework to demonstrate its practicality and effectiveness in two real-world scenarios: (1) human-bot interaction and (2) dataset augmentation. To the best of our knowledge, this is the first study to assess the *image-sharing* ability of various LLMs in a zero-shot setting. We make our source code and dataset publicly available [1].

Figure 1: An illustration of human's internal two-stage system of the image-sharing behavior.

1 Introduction

People often share a variety of images during interactions via instant messaging tools. In practice theory, this is referred to as *photo-sharing* behavior (Lobinger, 2016), which is interpreted as a communicative practice. [2] This behavior operates internally through a two-stage system (Zang et al., 2021): (1) *when to share* and (2) *what to share*. For example, as shown in Figure 1, we first discern the appropriate moment (i.e., decision) for sharing an image with certain intent based on one sentence that invokes the image-sharing behavior. Then, we share a relevant image at that moment, either by searching the internet or using photos taken on our mobile devices. This work primarily focuses on unlocking the image-sharing behavior capabilities of Large Language Models (LLMs) in a zero-shot manner.

However, recent studies related to multi-modal dialogue exhibit two significant limitations; **(1) Over-Simplification.** As shown in Figure 1, we humans decide to share images with an underlying intent, such as *visual clarification*, due to a certain triggering sentence (e.g., "*What did you eat?*") in the previous dialogue context. Nevertheless, existing studies (Zang et al., 2021; Feng et al., 2022) have reduced this complex behavior to a binary format (i.e., "yes" or "no"), which oversimplifies the complexity of image-sharing behavior. This sim-

[1] https://github.com/passing2961/DribeR

[2] From now on, we refer to this as *image-sharing* behavior, given that "image" is a broader concept than "photo," thereby providing more flexibility to language models.

Findings of the Association for Computational Linguistics: ACL 2024, pages 692–713
August 11-16, 2024

plification limits a comprehensive understanding of image-sharing behavior. (2) **Limited Understanding of Dialogue.** Since the key evidence is scattered across the entire dialogue context (Chae et al., 2023), understanding the linguistic cues underlying dialogue context is critical (Wang et al., 2023a) for retrieving images relevant to the dialogue. Nevertheless, existing systems (Lee et al., 2021; Zang et al., 2021; Feng et al., 2022) have primarily leveraged a dialogue-image matching score based on the cosine similarity between the two modalities (i.e., dialogue and image) to retrieve the relevant image at the image-sharing moment. This approach leads to a lack of capacity to comprehend the dialogue context due to the limitation of dual encoder structures (e.g., CLIP) in fully grasping the dialogue context (Yin et al., 2024).

This work explores whether LLMs contain the *image-sharing* capability, primarily focusing on a zero-shot performance. To this end, we introduce **D**ecide, Desc**ribe**, and **R**etrieve (DRIBER), a gradient-free, extensible, and generalizable framework to unlock this *image-sharing* capability of LLMs through in-context zero-shot learning. Broadly, DRIBER consists of three stages: (1) deciding the image-sharing behavior with the intent, (2) describing the image description relevant to the previous dialogue, and (3) retrieving the relevant image to the image description. The overall pipeline is illustrated in Figure 3. In addition, we introduce PHOTOCHAT++, an extended version of PHOTOCHAT. PHOTOCHAT is a multi-modal dialogue dataset constructed via crowdsourcing for human-human interaction. PHOTOCHAT++ contains six intent labels, a triggering sentence, and salient information (e.g., "words" or "phrases") to invoke the image-sharing behavior. In our experiments, we successfully unlock the *image-sharing* capability of LLMs in a zero-shot setting with the aid of DRIBER, with ChatGPT achieving state-of-the-art performance. Using the PHOTOCHAT++ dataset, we demonstrate that image-sharing is a challenging task for both humans and LLMs even if we apply the few-shot setting and Chain-of-Thought reasoning. Our extensive experiments further confirm that our framework is effective and versatile in real-world applications, specifically in (1) human-bot interaction dialogue and (2) dataset augmentation.

In summary, our main contributions are as follows: 1) We introduce the Decide, Describe, Retrieve (DRIBER) framework, designed to evaluate the *image-sharing* ability of LLMs in a zero-shot setting. 2) For a comprehensive assessment of LLMs' image-sharing capabilities, we present the PHOTOCHAT dataset, enriched with additional information such as intent, triggering sentences, image descriptions, and salient information. 3) Compared to the existing method, Experimental results show that DRIBER with LLMs achieves competitive zero-shot performance, even without additional training. 4) To the best of our knowledge, this is the first study to test the *image-sharing* capability of LLMs through zero-shot prompting.

2 Overview of PHOTOCHAT++

The PHOTOCHAT++ provides additional information (i.e., intent, triggering sentence, image description, and salient information) related to the image-sharing behavior by annotating the original PHOTOCHAT test set. The purpose of this dataset is to thoroughly assess the image-sharing capability of LLMs based on the internal operation system of humans. We describe the details of the human annotation procedure in Appendix E.

2.1 Preliminary: PHOTOCHAT

The PHOTOCHAT (Zang et al., 2021) is a human-human multi-modal dialogue dataset constructed through crowdsourcing. This dataset contains 10k multi-modal dialogues, where each dialogue $\mathcal{D} = \{(u_1, s_1), ..., (u_{t-1}, s_{t-1}), (i_t, s_t), (u_{t+1}, s_{t+1}), ..., (u_N, s_N)\}$ in the dataset contains only one image i_t to be shared at turn t. The N and $s_j \in \{0, 1\}$ denote the number of dialogue turns and speaker information, respectively. In addition, they define two tasks by decomposing the image-sharing behavior — a photo-sharing intent prediction task and an image retrieval task. The formulations are described as follows.

Task 1: Photo-Sharing Decision Prediction. Given the dialogue history $(u_j)_1^{t-1}$ and the corresponding speaker information $(s_j)_1^{t-1}$, this task aims to predict whether it is appropriate to share the image at turn t in the binary classification formulation, where the label $y \in \{0, 1\}$. [3]

Task 2: Image Retrieval. Given the dialogue history $(u_j)_1^{t-1}$ and the corresponding speaker information $(s_j)_1^{t-1}$, this task aims to retrieve most

[3]Originally, this task is called as "photo-sharing intent prediction". However, we consider that the "decision" term is more suitable than the term "intent" in this task because this task just predicts "*yes*" or "*no*".

Intent	Purpose
Information Dissemination	To convey important information
Social Bonding	To strengthen social relationship
Humor and Entertainment	To amuse or entertain
Visual Clarification	To clarify complex situations
Topic Transition	To change the topic of dialogue
Expression of Emotion or Opinion	To express emotions, opinions, or reactions

Table 1: Intent Category for Image-Sharing Behavior.

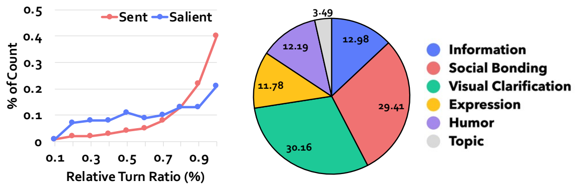

Figure 2: Analysis of PHOTOCHAT++. (Left) the distribution of triggering sentence and salient information. (Right) the intent distribution.

appropriate image at turn t from the image candidate set.

2.2 Intent Category

As shown in Table 1, we design six intent labels for image-sharing behavior: Information Dissemination, Social Bonding, Humor and Entertainment, Visual Clarification, Topic Transition, and Expression of Emotions or Opinions. The detailed explanation is described in Appendix B.

2.3 Collecting Annotations from Humans

We collect additional information through the human annotation process based on the following considerations. **(1) Intent.** Recognizing that the predefined intents for image-sharing behavior are not always mutually exclusive and can intersect based on the context and content of the image, we instruct annotators to select all intents that are applicable. **(2) Triggering Sentence.** Annotators are asked to identify and highlight the most significant sentence (i.e., only one sentence) from the preceding dialogue that contributed to the decision to share an image. **(3) Image Description.** We request annotators to write only one image description, beginning with prefix phrases such as "An image of" or "A photo of". This format aims to standardize the descriptions for ease of analysis. **(4) Salient Information.** We ask annotators to highlight all words or phrases they focus on in generating the image description. This step is crucial for understanding the key elements that influence the description process in the human internal mind.

2.4 Analysis of PHOTOCHAT++

We analyze PHOTOCHAT++, focusing on intent distribution and the distribution of triggering sentences and salient information.

Intent Distribution. Visual clarification is the most prevalent, indicating that images are often used in social conversations to aid understanding. Social bonding was also common, likely because the PHOTOCHAT is designed for social bonding. The least common is Topic transition, which is logical since topic transitions are more likely in longer conversations. However, the PHOTOCHAT typically has only 12 utterances per conversation, shorter than many long-term dialogue datasets. Topic transition is interesting and related to proactiveness, suggesting that the future creation of long-term multi-modal dialogue datasets could be beneficial.

Distribution of Key Information. We analyze which utterances in the dialogue just before image sharing trigger this behavior and where in the previous dialogue people focus when creating image descriptions. It's evident that image-sharing behavior is often triggered by utterances immediately preceding the image-share. Notably, the words or phrases that people cognitively focus on when creating image descriptions are not only distributed in the immediate preceding utterance but also throughout the early and middle parts of the conversation. This shows the importance of a model's ability to understand the entire dialogue when performing dialogue-to-image retrieval.

3 Methodology

3.1 Input Prompt Template

To elicit the *image-sharing* ability of LLM in a zero-shot setting, we manually construct a QA-style prompt template for our framework. The prompt template consists of four main parts: `[instruction]`, `[dialogue]`, `[restrictions]`, and `[question]`. For each stage, we use different sentences for `[instruction]`, `[restrictions]`, and `[question]`. To explicitly control the model's output, we add `[restrictions]` with `Restrictions:`. In addition, to make the `[dialogue]` more natural, we replace s_j with Top-1K common names of US SSN applicants from 1990 to 2021 [4], followed by a previous work (Kim

[4] https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-data

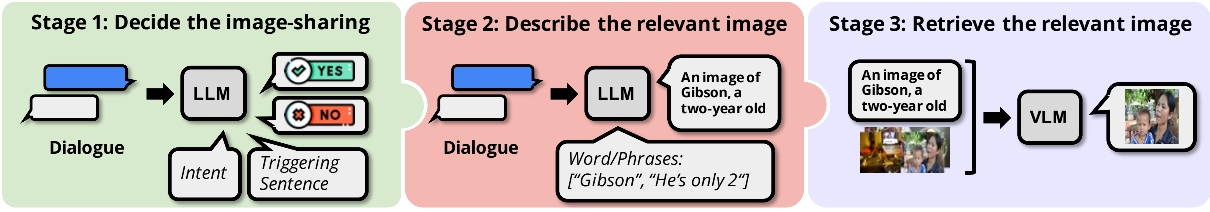

Figure 3: An illustration of our proposed framework: **Decide**, **Describe**, and **Retrieve** (DRIBER)

et al., 2022). We present the prompt template in Appendix.

3.2 Stage 1: Deciding Image-Sharing

Given the dialogue history $(u_j)_1^{t-1}$, the corresponding speaker information $(s_j)_1^{t-1}$, prompt p_1 and model $\mathcal{M}$, this stage generates an output $\hat{y_o}$ about whether it is appropriate to share the image: $\hat{y_o} = \mathcal{M}(p_1)$. Previous studies (Zang et al., 2021; Li et al., 2023b) ask the model $\mathcal{M}$ to predict the decision, through a simple binary response (*yes* or *no*), whether sharing an image is proper or not. In this work, we consider additional information, *intent* and *triggering sentence*, to investigate the image-sharing capability of the model $\mathcal{M}$ thoroughly. Specifically, the model $\mathcal{M}$ predicts all possible intents among the multiple-choice options and generates one sentence that invokes the image-sharing behavior for the same dialogue. When the predicted decision is "*yes*", then we generate the relevant image description for the same dialogue in stage 2.

3.3 Stage 2: Describing Relevant Image

Next, stage 2 uses the same model $\mathcal{M}$ to generate the image description $\hat{y}_{\text{desc}}$ conditioned on the previous dialogue history $(u_j)_1^{t-1}$: $\hat{y}_{\text{desc}} = \mathcal{M}(p_2)$. Then, instead of using the dialogue $(u_j)_1^{t-1}$ to retrieve the image, as in previous works, we use $\hat{y}_{\text{desc}}$ as a query to retrieve the relevant image. Therefore, the result of retrieved image is bound to depend on the quality of the corresponding $\hat{y}_{\text{desc}}$.

3.4 Stage 3: Retrieving Relevant Image

This stage aims to retrieve the relevant image based on the generated image description ($\hat{y}_{\text{desc}}$) from stage 2 by leveraging the vision-and-language pre-trained model (VLM). In this work, we use VLM as CLIP (Radford et al., 2021), which is a well-generalized and widely used model in various multi-modal tasks. However, our framework could work with any VLMs, such as BLIP (Li et al., 2022) or ALIGN (Jia et al., 2021).

4 Evaluating Image-Sharing Capability

To verify the performance for each stage, we measure the performance on various evaluation metrics. Each metrics are described as follows.

4.1 Evaluation for Stage 1

To understand how well the model $\mathcal{M}$ predicts the decision, intent, and triggering sentence, we report three types of metrics:

(1) DECISION$_{\text{[Y/N]}}$. The model should predict the decision with "*yes*" or "*no*". We measure the macro F1 score between the ground-truth decision label y_{D} and the predicted one $\hat{y}_{\text{D}}$.

(2) INTENT$_{\text{[CHOICE]}}$. The model should choose all possible intents among the multiple-choice options. Given the ground-truth intents y_{I} and the predicted intents $\hat{y}_{\text{I}}$, we measure the F1 score between y_{I} and $\hat{y}_{\text{I}}$.

(3) SENTENCE$_{\text{[DIST.]}}$. The model should generate one most contributed sentence to the image-sharing behavior in a free-form response. We measure the distance between the ground-truth sentence y_{S} (in PHOTOCHAT++) and the predicted response $\hat{y}_{\text{S}}$ by using the token F1 score. Specifically, we get the averaged token F1 score between $\hat{y}_{\text{S}}$ and the ground-truth sentences in PHOTOCHAT++.

4.2 Evaluation for Stage 2

Based on the theoretical view (Jaimes and Chang, 1999; Santurkar et al., 2022), we evaluate the image description in terms of descriptiveness and completeness. In addition, we evaluate whether the model has a similar cognitive process to a human with respect to salient information. We report three types of metrics:

(1) DESCRIPTIVENESS (Santurkar et al., 2022). This measures the inter-modal consistency: how much the image description ($\hat{y}_{\text{desc}}$) can replace the ground-truth image provided by PHOTOCHAT. However, measuring the descriptiveness is infeasible. Thus, previous work (Santurkar et al., 2022) approximates it with the help of an image-text matching model (e.g., BLIP (Li et al., 2023a)). Here, we use CLIPScore (Hessel et al., 2021).

(2) COMPLETENESS (Santurkar et al., 2022). This involves how much the generated image description ($\hat{y}_{\text{desc}}$) represents the object in the image. We measure the intersection ratio between the object $\hat{y}_{\text{obj}}$ in $\hat{y}_{\text{desc}}$ and the object y_{obj} in the ground-truth image is $\frac{|y_{\text{obj}} \cap \hat{y}_{\text{obj}}|}{|y_{\text{obj}}|}$. While y_{obj} is provided by the PHOTOCHAT, we extract the object $\hat{y}_{\text{obj}}$ in y_{desc} by prompting ChatGPT to extract the object in the given object categories from PHOTOCHAT. The prompt we used is presented in Appendix.

(3) CONSISTENCY. This measures the intra-modal consistency between the generated image description ($\hat{y}_{\text{desc}}$) and the human-written image description in PHOTOCHAT++. We measure the averaged sentence similarity using Sentence-BERT (Reimers and Gurevych, 2019) [5].

(4) SALIENT INFORMATION. To understand whether LLMs show similar cognitive processes when they generate image descriptions, we ask the model $\mathcal{M}$ to generate the salient words or phrases that are used to generate the image description. We measure the average token F1 score between the model prediction and ground-truth salient information in PHOTOCHAT++.

4.3 Evaluation for Stage 3

We evaluate how the generated image description retrieves relevant images using the CLIP model. We use a standard metric Recall@$\{1, 5, 10\}$ and mean reciprocal rank (MRR).

5 Experiments

5.1 Large Language Models in DRIBER

The primary objective is to assess the *image-sharing* capability of DRIBER equipped with LLM in terms of zero-shot performance, which necessitates complex reasoning. To achieve this, it is inevitable to leverage instruction-tuned large language models. For proprietary LLMs, we evaluate 2 models in total: 1) ChatGPT (OpenAI, 2023a), and 2) GPT-4 (OpenAI, 2023b). [6]. For open-sourced LLMs, we evaluate 3 models in total: 1) VICUNA 13B (Chiang et al., 2023), 2) FALCON INSTRUCT (40B; (Wang et al., 2023c)), and 3) LLAMA2 CHAT 70B (Touvron et al., 2023). We present the hyperparameter settings for each stage in Appendix J.

[5] We use sentence-transformers/all-roberta-large-v1 model.

[6] We conduct experiments with all language models by calling the OpenAI API between November-2023 and December-2023.

Model	F1	Precision	Recall
Fine-tuned Model			
ALBERT-base	52.2	44.8	62.7
BERT-base	53.2	56.1	50.6
T5-base	58.1	58.2	57.9
T5-3B	58.9	54.1	64.6
ViLT	52.4	55.4	58.9
PaCE	63.8	63.3	68.0
Ours			
DRIBER ChatGPT 0613	**65.6**	**66.7**	64.7
DRIBER ChatGPT 1106	64.0	63.1	65.1
DRIBER GPT-4 1106	45.6	56.5	**68.1**
DRIBER Vicuna-13B	65.5	64.3	67.1
DRIBER LLaMa2-Chat-70B	40.9	55.0	63.3
DRIBER Falcon-Instruct-40B	62.8	62.4	63.3

Table 2: Zero-shot results of DECISION$_{\text{[Y/N]}}$. The performance of fine-tuned models is reported from the previous work (Li et al., 2023b).

Model	F1	Precision	Recall	Refusal Ratio
DRIBER ChatGPT 1106 (0 shot)	**64.0**	**63.1**	**65.1**	0.0
DRIBER ChatGPT 1106 (1 shot)	60.8	60.5	61.0	0.1
DRIBER ChatGPT 1106 (2 shot)	37.7	53.8	59.4	4.2
DRIBER ChatGPT 1106 (4 shot)	27.2	53.5	55.5	10.6
DRIBER ChatGPT 1106 (8 shot)	26.5	53.7	54.8	**20.4**

Table 3: Few-shot results of DECISION$_{\text{[Y/N]}}$ in DRIBER, combined with ChatGPT 1106, by varying the number of few-shot examples. The Refusal Ratio (%) indicates the ratio of generated answers where DRIBER refuses to provide the decision.

5.2 Results

DRIBER unlocks the image-sharing capability. Table 2 shows the zero-shot results of various models of DECISION$_{\text{[Y/N]}}$ on PHOTOCHAT. Compared to the fine-tuned models, we mainly show that most LLMs can share images by understanding the given dialogue context without additional training on PHOTOCHAT. This is notable when compared to the performance of models that have been fine-tuned. Such a finding suggests that DRIBER can be effectively utilized in social dialogues requiring an understanding of, and an ability to imagine, interactions between multiple individuals, which is benefited from the power of instruction-tuned LLM. Notably, DRIBER with ChatGPT 0613 establishes a new state-of-the-art performance on DECISION$_{\text{[Y/N]}}$, surpassing the PaCE (Li et al., 2023b). Interestingly, DRIBER with GPT-4 1106 and LLaMa-2-Chat 70B exhibit lower performance, highlighting that the task of image sharing remains a challenging one for these models.

Model	F1	Precision	Recall	FN	FP
DRIBER ChatGPT 0613	65.6	66.7	64.7	629	499
DRIBER ChatGPT 0613 + CoT	65.1	65.1	65.2	606	617
DRIBER ChatGPT 1106	64.0	63.1	65.1	593	758
DRIBER ChatGPT 1106 + CoT	47.6	53.8	60.7	359	3648

Table 4: Zero-shot results of $\text{DECISION}_{[Y/N]}$ in DRIBER, combined with two different versions of ChatGPT when CoT is applied.

Few-shot examples confuse DRIBER in deciding the image-sharing behavior. As shown in Table 3, we evaluate the few-shot performance of DRIBER with ChatGPT 1106 by varying the number of few-shot examples. Interestingly, increasing the number of few-shot examples does not improve but rather decreases the $\text{DECISION}_{[Y/N]}$ performance. Specifically, there is an inverse relationship between the number of few-shot examples and the performance of $\text{DECISION}_{[Y/N]}$. This finding contrasts with previous research on factual reasoning, such as in (Kojima et al., 2022), where few-shot examples are shown to improve the capabilities of LLMs. To thoroughly investigate this phenomenon, we measured the ratio of instances (e.g., "I'm sorry, I cannot assist with that request") where DRIBER refuses to make a decision (Refusal Ratio). As indicated in Table 3, the refusal ratio significantly increases as performance decreases. This indicates that the presence of diverse decisions within few-shot examples confuses DRIBER in making image-sharing decisions.

Chain-of-Thought is not effective in the image-sharing capability. Table 4 shows the zero-shot performance of $\text{DECISION}_{[Y/N]}$ when we apply Chain-of-Thought (CoT) reasoning, specifically using the prompt "Let's think step by step.", as followed by (Kojima et al., 2022). Generally, we observe that CoT does not enhance the performance of $\text{DECISION}_{[Y/N]}$. This suggests that inducing the model to think about the possibility of image-sharing behavior sequentially often leads to a conclusion in favor of sharing images. In fact, it is noted that the proportion of FP increases. These results confirm that the image-sharing behavior has significant subject property.

DRIBER enhances the dialogue-to-image retrieval task. Table 5 presents the zero-shot results of the dialogue-to-image retrieval task. Notably, DRIBER, combined with various LLMs, significantly outperforms comparative models by a large margin of approximately 7.1% (an absolute value). In addition, we observe that even large-scale pre-trained vision-and-language models (ViLT, PaCE, CLIP) underperform compared to human performance. This suggests that these models have limitations in comprehending dialogues and summarizing relevant image descriptions accurately. The image-sharing behavior also presents a considerable challenge to humans, underscoring its complexity.

Model	R@1	R@5	R@10	MRR
Human	19.6	37.5	44.7	-
Fine-tuned Performance				
BM25	6.6	15.4	23.0	-
DE	9.0	26.4	35.7	-
VSE++	10.2	25.4	34.2	-
SCAN	10.4	27.0	37.1	-
VLMo	13.8	30.0	39.4	-
ViLT	11.5	25.6	33.8	-
PaCE	15.2	36.7	49.6	-
DialCLIP	19.5	44.0	55.8	-
VLM, zero-shot				
CLIP-base	13.7	28.0	35.2	20.8
CLIP-large	14.1	28.7	35.3	21.5
Large Multi-Modal Model				
LLaVA v1.5 7B	11.1	26.5	33.3	18.8
LLaVA v1.5 13B	12.1	25.6	32.3	19.3
MiniGPT-4 Vicuna 7B	11.6	26.5	34.0	19.1
MiniGPT-4 Vicuna 13B	11.7	27.7	35.5	19.8
Qwen-VL-Chat 7B	12.1	27.4	36.1	20.2
GPT4-V	13.8	27.9	35.9	21.3
Ours				
DRIBER ChatGPT 0613	26.6	46.1	54.2	36.0
DRIBER ChatGPT 1106	26.3	45.6	**54.3**	35.4
DRIBER GPT-4 1106	**28.3**	**47.4**	**55.2**	**37.6**
DRIBER Vicuna-13B	25.8	45.0	53.1	35.0
DRIBER LLaMa2-Chat-70B	24.5	43.5	52.6	34.0

Table 5: Zero-shot results of DRIBER with various LLMs on the image retrieval task. The performance of fine-tuned models is reported from the previous work (Li et al., 2023b).

Furthermore, we assess the image-sharing capability of recent large multi-modal models such as LLaVA v1.5, MiniGPT-4, and GPT4-V, which show remarkable performance on various visual-grounded language tasks (e.g., VQA). Given that these models are not specifically designed for cross-modal retrieval tasks, we evaluate them by measuring the text similarity score between the generated image descriptions and dialogue history, using Sentence-BERT (Reimers and Gurevych, 2019), as proposed in prior works. Among these models, GPT4-V achieves the best performance. However, their performance still falls short of that achieved by CLIP-large. This result indicates that, despite their versatility in visual-grounded language tasks, these models are not the optimal solution for tasks involving image-sharing behavior, likely due to their structural limitations.

Model	Descriptiveness	Completeness	Consistency	Salient
Human	0.1895	15.62	-	-
DRIBER$_{\text{ChatGPT 0613}}$	0.1846	19.94	**0.62**	**0.1288**
DRIBER$_{\text{ChatGPT 1106}}$	0.1873	19.48	**0.62**	0.0836
DRIBER$_{\text{GPT-4 1106}}$	**0.1988**	**20.80**	0.61	0.1013

Table 6: Zero-shot results of stage 2.

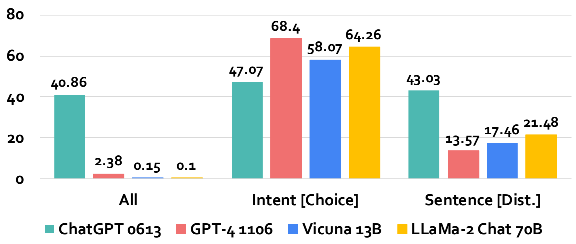

Figure 4: Results for INTENT$_{[\text{CHOICE}]}$ and SENTENCE$_{[\text{DIST.}]}$ are presented, with ALL signifying instances where DRIBER correctly identifies DECISION$_{[\text{Y/N}]}$, INTENT$_{[\text{CHOICE}]}$, and SENTENCE$_{[\text{DIST.}]}$ simultaneously.

DRIBER has visual imagination ability better than humans. Table 6 summarizes the zero-shot results from stage 2, focusing on DESCRIPTIVENESS, COMPLETENESS, CONSISTENCY, and SALIENT INFORMATION. Contrasting with the DECISION$_{[\text{Y/N}]}$ results, DRIBER with GPT-4 1106 outperforms DRIBER (w/ ChatGPT) in describing images, even surpassing human performance in DESCRIPTIVENESS and COMPLETENESS. These findings support the notion that DRIBER$_{\text{GPT-4 1106}}$ possesses notable visual imagination capabilities (Lu et al., 2022; Zhu et al., 2023b; Lu et al., 2023). However, DRIBER$_{\text{ChatGPT 0613}}$ demonstrates superior attention to SALIENT INFORMATION compared to DRIBER$_{\text{GPT-4 1106}}$. This suggests that the task of image sharing remains a significant challenge for these models.

Zero-shot Results of INTENT$_{[\text{CHOICE}]}$. Figure 4 shows the zero-shot results of INTENT$_{[\text{CHOICE}]}$ and SENTENCE$_{[\text{DIST.}]}$ on PHOTOCHAT++. Our findings reveal that DRIBER with ChatGPT 0613 exhibits a lower performance in INTENT$_{[\text{CHOICE}]}$ than other models. However, regarding performance across ALL categories, DRIBER with ChatGPT 0613 significantly outperforms the others. This observed decrease in ALL category performance among the other models indicates a selective impact on varying aspects of DECISION$_{[\text{Y/N}]}$, INTENT$_{[\text{CHOICE}]}$, and SENTENCE$_{[\text{DIST.}]}$.

5.3 Applications

We demonstrate the applicability and extensibility of our pipeline in two distinct applications: (1) Human-Bot Interaction, and (2) Dataset Augmentation.

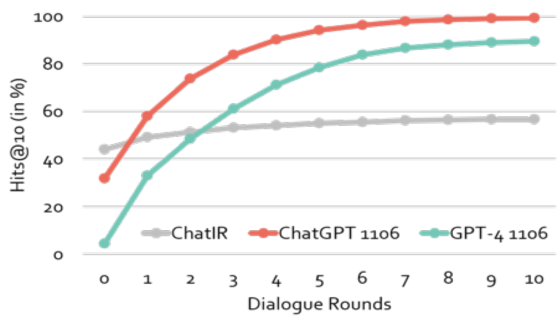

Figure 5: Zero-shot results of Hits@10 (%) across multiple dialogue rounds are presented, comparing the ChatIR system (Levy et al., 2023) with DRIBER. It should be noted that a dialogue comprising 0 rounds indicates that only the image description is provided to the model.

Application 1: Human-Bot Interaction. To validate the effectiveness of our pipeline in practical scenarios, we evaluate our pipeline on real human-bot interaction datasets based on VisDial (Das et al., 2017) dataset, which is introduced in the prior work (Levy et al., 2023). This specific scenario does not necessitate the decision-making process regarding image sharing. Therefore, DRIBER is tailored to include only stages 2 and 3, showcasing its flexibility. As shown in Figure 5, DRIBER significantly outperforms the recent ChatIR system (Levy et al., 2023), indicating its strong generalization performance. In addition, we find that the performance of DRIBER is enhanced when provided with a more extensive dialogue context. These results indicate a strong and robust understanding of the dialogue context within DRIBER, benefiting from the enormous ability of LLM.

Application 2: Dataset Augmentation. Previous studies (Lee et al., 2021; Kim et al., 2022; Chun et al., 2022) have reported that machine-annotated or generated datasets significantly enhance generalization performance. Drawing inspiration from these works, we augment PHOTOCHAT with DRIBER equipped with ChatGPT 1106 by providing the full dialogue context to the LLM. The motivation for this is that given that the dialogue context in PHOTOCHAT is already fixed, it is crucial to find image-sharing moments that wouldn't disrupt the existing conversation flow after inserting relevant images at the image-sharing turns. Thus, we slightly modify the input prompt, as detailed in Appendix. Additionally, we prompt

Eval →	PhotoChat			MMDialog		
Train ↓	R@1	R@5	R@10	R@1	R@5	R@10
Image Retrieval						
PhotoChat	$16.51_{\pm 0.27}$	$43.37_{\pm 0.82}$	$60.44_{\pm 1.59}$	$5.88_{\pm 0.21}$	$19.21_{\pm 0.92}$	$29.95_{\pm 1.3}$
aug-PhotoChat	$\mathbf{16.92}_{\pm 1.18}$	$\mathbf{44.89}_{\pm 1.18}$	$\mathbf{61.77}_{\pm 1.05}$	$\mathbf{8.25}_{\pm 0.47}$	$\mathbf{24.95}_{\pm 1.21}$	$\mathbf{37.0}_{\pm 1.65}$
Next Response Prediction						
PhotoChat	$6.02_{\pm 0.26}$	$19.81_{\pm 0.78}$	$31.67_{\pm 1.68}$	$2.34_{\pm 0.24}$	$9.27_{\pm 0.89}$	$16.22_{\pm 1.25}$
aug-PhotoChat	$\mathbf{6.43}_{\pm 0.93}$	$\mathbf{23.06}_{\pm 1.52}$	$\mathbf{34.24}_{\pm 1.15}$	$\mathbf{2.67}_{\pm 0.07}$	$\mathbf{9.86}_{\pm 0.26}$	$\mathbf{16.29}_{\pm 0.38}$

Table 7: We report the text and image retrieval performance across five runs on three multi-modal dialogue datasets: PhotoChat and MMDialog (Feng et al., 2022).

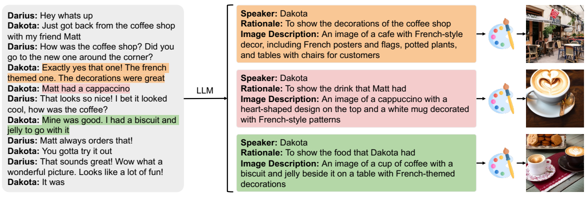

Figure 6: **An example of *aug*-PhotoChat .** Followed by our pipeline, we construct the *aug*-PhotoChat by generating appropriate images using Stable Diffusion ( ) by prompting predicted image descriptions from LLM (i.e., ChatGPT 1106), which are highlighted in orange, pink, and green boxes.

the LLM to generate a rationale for the act of image sharing, thereby gaining insight into the LLM's judgment. Once an image-sharing moment is identified, we align a relevant image to this moment using Stable Diffusion (Rombach et al., 2022)[7]. This step is crucial as the LLM sometimes overgenerates image descriptions, making it challenging for CLIP to align the relevant image using existing source image datasets like Conceptual Captions 3M (Sharma et al., 2018). We refer to this augmented dataset as *aug*-PhotoChat .

To assess if *aug*-PhotoChat improves generalization performance on unseen multi-modal dialogue datasets, we implement simple text and image retrieval models (details in Appendix K). We train these models on both PhotoChat and *aug*-PhotoChat and evaluate them on PhotoChat and MMDialog (Feng et al., 2022), a million-scale multi-modal dialogue dataset. Table 7 demonstrates that models trained on *aug*-PhotoChat outperform those trained on the two multi-modal dialogue datasets. Notably, there is a significant performance boost on MMDialog with *aug*-PhotoChat , highlighting the effectiveness of our pipeline in constructing *aug*-PhotoChat . An example from *aug*-PhotoChat is presented in Figure 6.

[7] https://huggingface.co/stabilityai/stable-diffusion-2-1-base

6 Related Work

Multi-Modal Dialogue Dataset. Existing studies predominantly fall into two categories, depending on whether the image in the dialogue is *grounded* or *sharing*. Image-grounded dialogue tasks are designed to answer questions (Antol et al., 2015; Das et al., 2017; Kottur et al., 2019) or generate natural conversations (Shuster et al., 2018; Meng et al., 2020; Wang et al., 2021b; Zheng et al., 2021) about given images. Nevertheless, it is common to share images pertinent to dialogue contexts in everyday conversations for the purpose of reinforcing social bonding, as well as enhancing engagement and interest. Recent studies have proposed datasets that encapsulate this image-sharing behavior. This has been achieved by collecting a human-human dialogue dataset (PhotoChat) via crowdsourcing (Zang et al., 2021), a large-scale dataset (MMDialog) from social media (Feng et al., 2022), or constructing datasets automatically using vision-and-language models (Lee et al., 2021). In this work, our focus is exclusively on the PhotoChat dataset to gain a deeper understanding of the *image-sharing* capabilities of LLMs. We do not include automatically constructed datasets or the MMDialog due to the considerable expense associated with conducting experiments using LLMs.

Prompting Large Language Models. Recent studies have witnessed the success of large language models, such as Instruct GPT-3 (Ouyang et al., 2022), ChatGPT (OpenAI, 2023a), GPT-4 (OpenAI, 2023b), in a zero-/few-shot performance. The use of these models, in conjunction with "prompt engineering," has unlocked the abilities of LLMs, even emergent ones (Wei et al., 2022a), across various tasks. These tasks range from dialogues (Lee et al., 2022a; Kim et al., 2022; Lee et al., 2022b), complex reasoning tasks (Wei et al., 2022b; Kojima et al., 2022), and theory-of-mind (Sap et al., 2022; Kosinski, 2023), to image classification(Yang et al., 2022; Pratt et al., 2022; Menon and Vondrick, 2022; Zhang et al., 2023) and multi-modality (Lu et al., 2023; Han et al., 2023).

7 Conclusion

In this paper, we explore the *image-sharing* capabilities of LLMs in a zero-shot prompting by introducing a three-stage framework: Decide, Describe, Retrieve (DribeR). Our extensive experiments demonstrate the effectiveness of our framework in enhancing zero-shot performance across both stages, with ChatGPT achieving state-of-the-art performance. We also reveal that the *image-sharing* ability is an emergent ability in the zero-shot prompting. With extensive experiments, we reveal the effectiveness of our framework on two useful applications. In future works, we will assess this capability in a few-shot setting using additional multi-modal dialogue datasets. We will construct a personalized multi-modal dialogue dataset using our framework.

Limitations

Here, we highlight some limitations of our work. Firstly, our prompt template is rather lengthy, which complicates expansion into the few-shot setting. We anticipate that conducting few-shot prompting to utilize the image-sharing capability of LLMs would result in better performance compared to zero-shot prompting. Secondly, LLMs tend to over-generate image descriptions even in the absence of specific demographic information such as age or appearance. For instance, in the description, "An image of a woman with long, brown hair wearing a flowy white dress and brown boots," there is no reference to long hair in the given dialogue. Providing additional information (e.g., persona) can enhance the relevance of image descriptions generated by LLMs.

Ethical Considerations

We report several potential issues with our proposed framework. First, generated image descriptions may propagate social bias because GPT-3 can still produce harmful content, including social bias and offensiveness(Baheti et al., 2021; Hartvigsen et al., 2022). Second, this issue has resulted in the inclusion of problematic descriptions in the constructed *aug*-PHOTOCHAT dataset, leading to socially-biased images generated using Stable Diffusion (Rombach et al., 2022). As a result, when vision-and-language models like CLIP (Radford et al., 2021) and DALL-E (Ramesh et al., 2021) are trained on this augmented dataset, they may exhibit social biases. As reported in (Wang et al., 2021a), even if we give the gender-neutral query to CLIP (Radford et al., 2021) model, the CLIP model sometimes retrieves images causing gender-bias issues. We are concerned that this problematic issue may exist in the augmented dataset. Therefore, the image retrieval model trained on this dataset may sometimes retrieve biased images. In addition, text-to-image generative models learn social biases from the augmented dataset, as reported in the prior work (Cho et al., 2022). We should consider this problem important when building a multimodal search model.

Acknowledgement

This work was supported by a grant of the KAIST-KT joint research project through AI Tech Lab, Institute of convergence Technology, funded by KT [Project No. G01230605, Development of Task-oriented Persona-based Dialogue Generation Combining Multi-modal Interaction and Knowledge Modeling].

References

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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. *arXiv preprint arXiv:2108.11830*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Hyungjoo Chae, Yongho Song, Kai Tzu-iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. *arXiv preprint arXiv:2310.09343*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*.

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. 2022. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 1–19. Springer.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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 multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.

Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. 2023. Champagne: Learning real-world conversation from large-scale web videos. *arXiv preprint arXiv:2303.09713*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Alejandro Jaimes and Shih-Fu Chang. 1999. Conceptual framework for indexing visual information at multiple levels. In *Internet Imaging*, volume 3964, pages 2–15. SPIE.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*.

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, Chae-Gyun Lim, and Ho-Jin Choi. 2022a. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022b. Personachatgen: Generating personalized dialogues using gpt-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. 2023. Chatting makes perfect–chat-based image retrieval. *arXiv preprint arXiv:2305.20062*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Yunshui Li, Binyuan Hui, ZhiChao Yin, Min Yang, Fei Huang, and Yongbin Li. 2023b. PaCE: Unified multi-modal dialogue pre-training with progressive and compositional experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13402–13416, Toronto, Canada. Association for Computational Linguistics.

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

Katharina Lobinger. 2016. Photographs as things–photographs of things. a texto-material perspective on photo-sharing practices. *Information, Communication & Society*, 19(4):475–488.

Yujie Lu, Pan Lu, Zhiyu Chen, Wanrong Zhu, Xin Eric Wang, and William Yang Wang. 2023. Multimodal procedural planning via dual text-image prompting. *arXiv preprint arXiv:2305.01795*.

Yujie Lu, Wanrong Zhu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Imagination-augmented natural language understanding. *arXiv preprint arXiv:2204.08535*.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*.

Sachit Menon and Carl Vondrick. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Icml*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023a. ChatGPT. `https://openai.com/blog/chatgpt/`.

OpenAI. 2023b. Gpt-4 technical report. *arXiv*.

OpenAI. 2023c. Gpt4-v. `https://openai.com/research/gpt-4v-system-card`.

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.

Sarah Pratt, Rosanne Liu, and Ali Farhadi. 2022. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*.

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.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. 2022. Is a caption worth a thousand images? a controlled study for representation learning. *arXiv preprint arXiv:2207.07635*.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023a. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021a. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*.

Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. 2021b. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *arXiv preprint arXiv:2109.12761*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023b. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023c. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2022. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *arXiv preprint arXiv:2211.11158*.

Zhichao Yin, Binyuan Hui, Min Yang, Fei Huang, and Yongbin Li. 2024. Dialclip: Empowering clip as multi-modal dialog retriever. *arXiv preprint arXiv:2401.01076*.

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*.

Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. 2023. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *arXiv preprint arXiv:2303.02151*.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Ke Lin. 2021. Mmchat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Wanrong Zhu, Xin Wang, An Yan, Miguel Eckstein, and William Yang Wang. 2023b. Imagine: An imagination-based automatic evaluation metric for natural language generation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 93–105.

A Additional Related Work

Large Multi-Modal Models. Recent studies have proposed large multi-modal models that demonstrate the surprising generalization performance on various visual-grounded language tasks, such as Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023), MiniGPT-4 (Zhu et al., 2023a), Qwen-VL (Bai et al., 2023), CogVLM (Wang et al., 2023b), and GPT4-V (OpenAI, 2023c). In this work, we assess the image-sharing capability of these models in a zero-shot setting.

B Details of PHOTOCHAT++

We describe each intent category as follows:

(1) Information Dissemination. This intent involves sharing images that convey important information, such as current news, economic updates, educational content, or plot summaries. These images are intended to inform or educate the recipient.

(2) Social Bonding. This category encompasses sharing personal photographs, memories of recent encounters, recollections of past events (i.e., memory recall), or images that connect to one's own experiences. Such image sharing is primarily used to strengthen social ties.

(3) Humor and Entertainment. This intent involves sharing images that aim to amuse or entertain, like humorous pictures or memes. The primary focus is on sharing light-hearted content to bring joy or laughter.

(4) Visual Clarification. Here, images serve as supplementary material to clarify complex concepts or situations, especially when textual conversation alone is insufficient. Images can significantly enhance understanding in such contexts.

(5) Topic Transition. In this category, images are utilized to change the topic of conversation or modify the mood subtly. This can be a strategic approach to redirecting a discussion or lightening the atmosphere.

(6) Expression of Emotion or Opinion. This intent includes the sharing of images to express emotions, opinions, or reactions that are challenging to communicate through text alone. It involves the use of emotive photography, art, or reaction images that effectively convey feelings or viewpoints.

C Comparison to Previous System

Figure 7 illustrates the overview of previous systems (Zang et al., 2021; Feng et al., 2022) and DRIBER.

D Additional Experiments

Figure 8 shows the zero-shot results of INTENT$_{[\text{CHOICE}]}$, SENTENCE$_{[\text{DIST.}]}$, and ALL when we apply zero-shot chain-of-thought (CoT) reasoning. For CoT, we follow (Kojima et al., 2022) and use the prompt "*Let's think step by step*". We observe that CoT does not improve the performance of INTENT$_{[\text{CHOICE}]}$, SENTENCE$_{[\text{DIST.}]}$, and ALL. These results suggest that CoT selectively improves the image-sharing capability of LLMs.

E Details of Human Annotation Procedure

E.1 Preparing Human Annotation

We first prepare 968 dialogues from the test set of PHOTOCHAT to collect additional information crucial for understanding image-sharing behavior in a manner akin to human comprehension. Then, we ask 12 human annotators to annotate various elements at the ground-truth moments of image-sharing in PHOTOCHAT. These elements include *intent*, *triggering sentence*, *salient information*, and *image description*. Notably, we intentionally do not show the ground-truth images from PHOTOCHAT during this annotation process. This process is adopted to explore the complexity and subjectivity inherent in image-sharing behavior.

F Human Evaluation Questionnaire

This section lists the questions and notes used for human evaluations.

F.1 Questions

- **What is the intent behind sharing the image?**

  Options: Information Dissemination / Social Bonding / Human and Entertainment / Visual / Clarification / Topic Transition / Expression of Emotion or Opinion

  Notes: You can select all of the following options that are possible. If the answer you have in mind is not listed, please write it in the space. If you write more than two sentences, then please divide them using the separator ";".

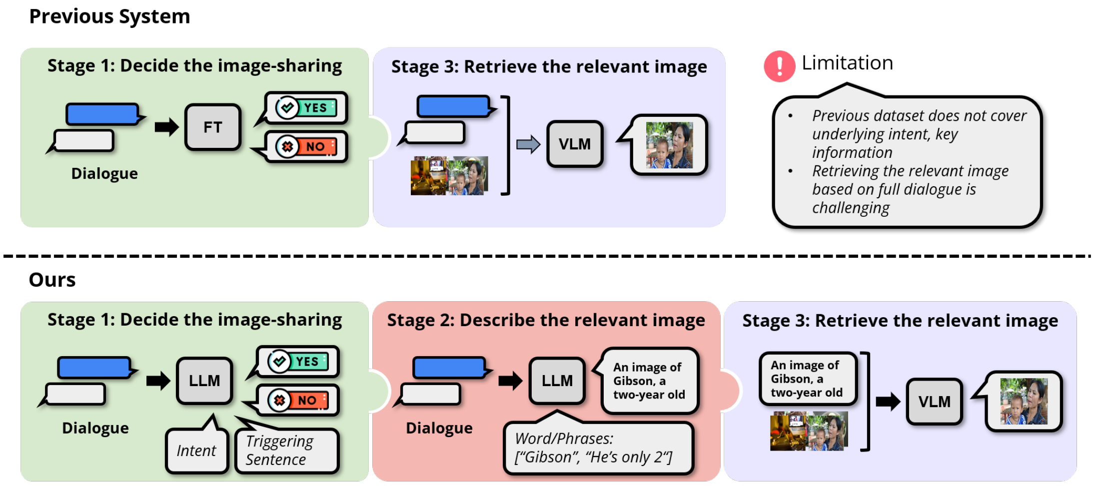

Figure 7: We compare DRIBER with the previous systems. FT denotes the fine-tuned model as shown in Table 2.

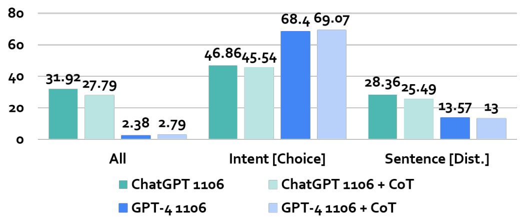

Figure 8: Results of INTENT[CHOICE], SENTENCE[DIST.], and ALL when CoT is applied.

- **Which sentence invokes the image-sharing behavior?**

  Notes: You should highlight the most contributed sentence (i.e., only one sentence) using "Sentence"

- **What kind of image would be appropriate to share at the [Sharing Image] turn?**

  Notes: You can write only one image description, starting with "An image of" or "A photo of".

- **What words or phrases do you focus on to write the image description?**

  Notes: You can highlight all words or phrases that you focus on using "Word/Phrase"

G Human Evaluation System

We show a screenshot of the human evaluation system in Figure . We implement this system using Label Studio (Tkachenko et al., 2020-2022).

H Details of Human Evaluation

We recruited 12 individuals, unknown to us, who are either graduate or undergraduate students. Prior

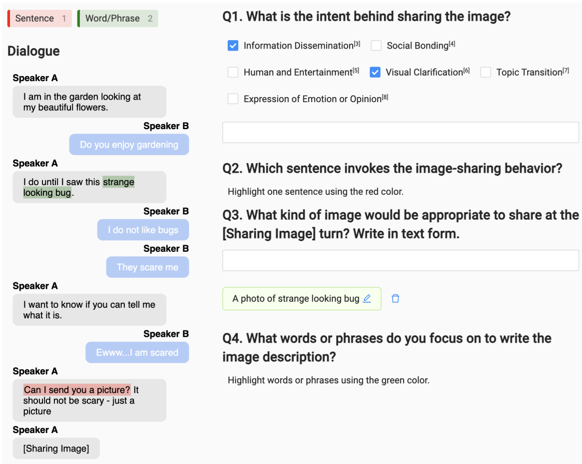

Figure 9: A screenshot of the human evaluation system for the human ratings.

to participating in the experiment, they were provided with comprehensive instruction on the task, an overview of the multi-modal dialogue dataset, and a detailed explanation of the evaluation criteria. This preparatory phase lasted approximately one hour.

I Discussions

Domain Generalizability. Towards Artificial General Intelligence (AGI), we need to show the proposed methods' generalization capability across different domains, such as medical imaging. We recognize the significance of this aspect and have designed our methodology to be both extensible and generalizable. Our approach's extensibility allows us to tailor DRIBER according to specific domains. For instance, in a specialized domain like medical, we could incorporate a domain-specific LLM, such as Med-PaLM 2 (Singhal et al., 2023b), or even GPT-4, which has demonstrated superior performance in tasks requiring medical knowledge (Nori et al., 2023), including outperforming domain-specific models like Med-PaLM (Singhal et al., 2023a) in the United States Medical Licensing Examination (USMLE). For the image retrieval aspect, a specialized model such as MedCLIP (Wang et al., 2022) could be employed to enhance retrieval accuracy in the medical context.

Towards Better Image-Sharing Ability. As shown in our experiment, the likelihood of performance improvement is high as the model's size increases or when it is trained with alignment to human preference. This suggests that the image-sharing ability is subjective and resembles human-like tasks. Therefore, receiving real-time feedback through interactive mode (a form of human-AI collaboration) and further training the model using the RLHF method could lead to better performance, aligning the model's actions favorably with image-sharing ability.

Furthermore, understanding conversational context is essential, and imbuing the model with the ability of perspective-taking, understanding situations from the user's point of view, could lead to performance enhancement. For instance, when a user is feeling down due to poor test results, the model could not only provide empathy through text but also share a picture of a dog based on the user's fondness for dogs and the current context of struggling with test scores, thereby offering multi-faceted empathy.

In addition, unlike image-grounded dialogue, image-sharing scenarios might lack explicit information from previous conversations. For instance, understanding what "it" refers to in "I love it" requires considering the preceding conversational context. Thus, it's important to consider coreference resolution. Moreover, while sharing images, incorporating information about significant utterances from previous dialogues or using keywords and keyphrases could likely improve performance.

As depicted in Figure 6's orange-generated results, the language model might sometimes over-generate due to excessive creativity. For instance, if the conversation only contains information about a coffee shop without mentioning "French-style," the model might still produce the word "French." Such cases could pose challenges in practical applications where inappropriate images could be retrieved.

In practical applications, it's beneficial to consider the user's satisfaction and share images that account for their personal information. For example, if a user mentions, "I work in a hotdog stand," and their friend, who also works there, has a picture related to selling hotdogs in their phone album, it would be more suitable to share an image depicting the user themselves selling hotdogs rather than an image with the friend. Of course, obtaining explicit consent for sharing personal information is crucial.

Additionally, beyond improving the image-sharing ability, at the application level, using videos could enhance user engagement. Exploring this avenue could be a promising direction for future research.

Intrinsic Properties of LLMs. We believe that the intrinsic properties of LLM, which have been experimentally proven in various studies, have influenced image-sharing ability.

- **Understanding the dialogue context:** It's essential to grasp the conversation topic holistically, emotional shifts between users, and general knowledge. Recent research results have shown that language models possess these abilities.

- **Understanding the interlocutor's mental state:** It is important to comprehend the interlocutor is situation to determine whether sharing an image is appropriate. For instance, if the interlocutor is upset, it might be better to respond empathetically rather than share an image. This ability is highly related to the Theory-of-Mind (ToM). Recently, LLMs have achieved competitive performance in Theory-of-Mind (ToM) tasks, which may influence image-sharing ability.

- **Understanding the intent:** From the model's perspective, sharing an image can be seen as intent. Many language models have demonstrated good performance in task-oriented dialogue tasks.

- **Visual imagination ability:** To share an appropriate image, one must imagine which image is best. This capability has been empirically proven in various recent studies. We investigated the C4 dataset, a representative pretraining dataset for LLMs, to analyze why this capability is manifested. The data discovered in C4 consists of pairs of images and their corresponding captions. These captions contain words/phrases related to visual imagination ability, such as "depict" and "photo of." Moreover, on blogs, images often appear consecutively along with stories. Due to these elements, the LLM learned an inherent visual, and imaginative capability during its pretraining phase.

J Implementation Details of LLMs

To evaluate the *image-sharing* capabilities of LLMs, we call ChatGPT (OpenAI, 2023a) and GPT-4 (OpenAI, 2023b) by calling OpenAI API. All experiments are conducted on two A100 (40GB) GPU. For each stage, the generation configuration is as follows:

- For Stage 1, we set maximum tokens to 1024, temperature to 0.0, frequency penalty to 0.0, presence penalty to 0.0, top_p to 1.0, and stop tokens to `\n\n`.
- For Stage 2, we set maximum tokens to 1024, temperature to 0.9, frequency penalty to 0.0, presence penalty to 0.4, top_p to 0.95, and stop tokens to a default setting.

K Details of Experimental Settings

To explore how our dataset affects both text and image retrieval tasks, we implement two simple and standard baseline retrieval models for text-to-image and image-to-text settings.

K.1 Task Definition

Follwing (Lee et al., 2021; Zang et al., 2021), we explain the formulation of two main tasks - next response prediction and image retrieval. Let us assume that we have a multi-modal dialogue $\mathcal{D} = \{(u_j, i_j, c_j)\}_1^N$ where N denotes the number of dialogue turns, and $j = t$ is the turn that an image sharing behavior occurs. Then, each task is formulated as follows.

Next response prediction is to predict the next utterance at turn $t+1$ given the dialogue history $(\{u_j\}_1^t)$ and image i_t.

Image retrieval is to retrieve relevant image at turn t given the dialogue history $(\{u_j\}_1^{t-1})$.

Following (Shuster et al., 2018; Lee et al., 2021), we set the number of retrieval candidates to 100 and use Recall@{1,5,10} and mean reciprocal rank (MRR) for the evaluation metrics.

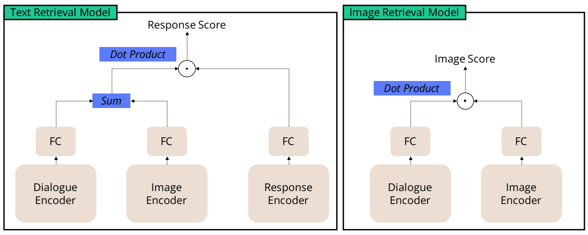

Figure 10: Architectures of two baseline models: Text retrieval and Image retrieval.

K.2 Baseline Models

As illustrated in Figure 10, we present the architecture of baseline models: the text retrieval and image retrieval models. We provide a detailed description of baseline models below.

Text Retrieval Model. The text retrieval model consists of three primary components: the dialogue encoder, the response encoder, and the image encoder. The dialogue encoder processes the entire dialogue history into a fixed-size representation using the BERT model (Devlin et al., 2018). The dialogue history includes up to three turns preceding the current turn, with each turn separated by the `[SEP]` token. The response encoder also converts responses into fixed-size representations, utilizing a different version of the BERT model than the dialogue encoder. After BERT processes the text, mean pooling is applied to the text representations for both encoders. These pooled representations then pass through a linear projection layer followed by the ReLU activation function (Nair and Hinton, 2010). The image encoder uses the CLIP-base model (Radford et al., 2021) to extract feature vectors from images. The dialogue and image feature vectors are then combined using element-wise addition. The loss is computed by taking the dot product between the response feature vector and the resulting summed vector.

Image Retrieval Model. The image retrieval model is composed of two main components: the dialogue encoder and the image encoder. The dialogue encoder uses the BERT-base model to convert the dialogue into a representation, followed by mean pooling of the text representations. For the image representation, the CLIP-base model is utilized. After encoding, the image and dialogue vectors are passed through their respective linear projection layers, each followed by a ReLU activation function. The loss is determined by calculating the dot product between the image feature vector and the dialogue vector.

K.3 Implementation Details

We implement baseline models based on PyTorch Lightning. All experiments are conducted on two A100 GPUs (40GB). To accelerate the training time, we apply distributed training to baselines. We follow the hyperparameter settings similar to the previous works (Lee et al., 2021; Zang et al., 2021), which are described as follows:

Text retrieval. In our experiment, we set the batch size to 256, the learning rate to 5e-5, and the gradient clipping value to 2.0. We use the AdamW optimizer with a cosine learning rate scheduler. We set the warm-up ratio as 0.1% and weight decay as 0.2.

Image retrieval. We set the batch size to 256. We also use the AdamW optimizer with an initial learning rate of 2e-5 and decaying 0.1%.

Training. Since our dataset contains several images per utterance, we randomly choose one image in each batch. We do not update the parameter of the image encoder.

L Rationale Distribution

We present the rationale distribution as shown in Table 8.

M Prompt Templates

Here, we present all prompt templates used in our work, such as restriction-based prompt templates for each stage, and several prompt templates for the ablation studies.

M.1 Prompt Templates

We present a prompt template for dataset augmentation in our proposed framework, as shown in Figure 11.

N More Examples of *aug*-PhotoChat

We provide more examples of *aug*-PhotoChat .

Verb	Object	Count	Example
provide	information	612	To provide more information about the moth she saw.
	context	445	To provide context for the conversation.
	representation	397	To provide a visual representation of the beverage person is talking about.
	evidence	215	To provide visual evidence of the fun time they had together.
show	interest	174	To show his interest in seeing the photo.
	image	173	To show the image of the letters he formed with the dough.
	person	149	To show person that he is okay with the weather.
	audience	111	To show the audience the fun person is having on his vacation.
share	image	145	To share the image of the birthday party.
	photo	13	To share the photo with person.
express	interest	30	To express interest in person's story
	reaction	27	To express her reaction to the image.
	excitement	17	To express excitement about the workshop.
	appreciation	12	To express her appreciation for the cake.
invite	person	38	To invite person to see the picture of the table.
ask	person	25	To ask person to share an image of his recent cooking.
	question	6	To ask a follow-up question about the image.
encourage	person	23	To encourage person to share his most memorable dinner.
introduce	image	12	To introduce the image.
	topic	8	To introduce the topic of the conversation.
gauge	interest	18	To gauge person's interest in the baked goods.
give	opportunity	13	To give person the opportunity to see a photo of Hannah.
engage	person	9	To engage person in the conversation and to show her the photo Zora sent.
emphasize	importance	8	To emphasize the importance of spending time with kids.
indicate	interest	7	To indicate person's interest in seeing the photo.

Table 8: **Rationale Distribution.** The top 20 most common root verbs and their up to 4 direct noun objects in the generated rationale. Only pairs with a count of 5 or more are included.

Prompt Template for Stage 1:
The following is a dialogue between `[speaker1]` and `[speaker2]`. You should share an image to make the following dialogue more interesting and engaging. The dialogue is provided line-by-line. In the given dialogue, select all utterances that are appropriate for sharing the image in the next turn, and write the speaker who will share the image after the selected utterance. You should also provide a rationale for your decision.

Dialogue:
`[dialogue]`

Restrictions:
(1) your answer should be in the format of "<UTTERANCE> | <SPEAKER> | <RATIONALE>".
(2) you MUST select the utterance in the given dialogue, NOT generate a new utterance.
(3) the rationale should be written starting with "To".

Answer:
1.

Prompt Template for Stage 2:
The following is a dialogue between `[speaker1]` and `[speaker2]`. The dialogue is provided line-by-line. `[speaker1]` shares an image in a given dialogue to make the following dialogue more interesting and engaging, marked in [Sharing Image]. Depict the most appropriate image to be shared in the next turn, in detail.

Dialogue:
`[dialogue]`

Restrictions:
(1) your answer should be written starting with "An image of" and in one sentence.
(2) you do NOT include the speaker's name (i.e., `[speaker1]`, `[speaker2]`) in the image description.
(3) you should share a realistic image, NOT memes.

Image Description:

Figure 11: **Prompt Templates for Dataset Augmentation.** A prompt template for stage 1 (**top**). A prompt template for stage 2 (**bottom**).

Prompt Template for Stage 1:
The following is a dialogue between `[speaker1]` and `[speaker2]`. The dialogue is provided line-by-line. You will be provided a list of the intent of sharing an image. In the given dialogue, you should predict whether it is appropriate for `[share_speaker]` to share an image in the next turn, the intent of the image-sharing behavior, and one sentence that invokes the image-sharing behavior.

List of Intent of Image-Sharing Behavior:
- Information Dissemination: This involves sharing images to communicate important information, such as news, economic updates, or educational material (infographic image), aiming to inform or educate
- Social Bonding: This involves sharing images to strengthen social connections, including personal photos and memories
- Humor and Entertainment: This involves sharing light-hearted images, such as funny pictures or memes, to entertain and bring joy
- Visual Clarification: This involves sharing images, such as diagrams, item-specific photos, or location images, to clarify complex concepts or situations
- Topic Transition: This involves sharing images to shift the conversation topic or mood
- Expression of Emotion or Opinion: This involves sharing images, such as emotive photos or art, to express emotions or opinions more effectively than text, succinctly conveying feelings or perspectives

Dialogue:
`[dialogue]`

Question: Is it appropriate for `[share_speaker]` to share an image in the next turn? If "Yes", choose all possible intents of sharing the image and provide only one sentence that invokes the image-sharing behavior.
Options:
(a) Information Dissemination
(b) Social Bonding
(c) Humor and Entertainment
(d) Visual Clarification
(e) Topic Transition
(f) Expression of Emotion or Opinion
Restrictions:
(1) You should provide your answer in a Python dictionary object with three keys, "Prediction", "Intent", and "Sentence".
(2) You should provide a binary answer (i.e., "yes" or "no") for the value of the "Prediction" key.
(3) You should choose all possible intents for the value of "Intent" key.
(4) You should provide the most contributed sentence (i.e., only one sentence) that invokes the image-sharing behavior for the value of "Sentence" key.
Answer:

Figure 12: **Prompt Templates for Stage 1.** A prompt template used in the ablation study in Stage 1.

Prompt Template for Stage 2:
The following is a dialogue between `[speaker1]` and `[speaker2]`. The dialogue is provided line-by-line. `[share_speaker]` shares an image in a given dialogue to make the following dialogue more interesting and engaging, as marked in [Sharing Image].

Dialogue:
`[dialogue]`

Question: What is the most appropriate image description to share in the [Sharing Image] turn?
Restrictions:
(1) You should provide your answer in a Python dictionary object with "Image Description" key.
Answer:

Figure 13: **Prompt Templates for Stage 2.** A prompt template used in the in Stage 2.

Prompt Template for One-Stage:
You will be provided a list of object categories and an image description. Your job is to detect the object in the image description and categorize the detected object into one of the categories in the list.

You must provide your answer in a Python dictionary object that has the category as the key and the corresponding object in the image description as the value.

Object Category List = ["Woman", "Man", "Girl", "Boy", "Human body", "Face", "Bagel", "Baked goods", "Beer", "Bread", "Burrito", "Cake", "Candy", "Cheese", "Cocktail", "Coffee", "Cookie", "Croissant", "Dessert", "Doughnut", "Drink", "Fast food", "French fries", "Hamburger", "Hot dog", "Ice cream", "Juice", "Milk", "Pancake", "Pasta", "Pizza", "Popcorn", "Salad", "Sandwich", "Seafood", "Snack", "Animal", "Alarm clock", "Backpack", "Blender", "Banjo", "Bed", "Belt", "Computer keyboard", "Computer mouse", "Curtain", "Guitar", "Hair dryer", "Hair spray", "Harmonica", "Humidifier", "Jacket", "Jeans", "Dress", "Earrings", "Necklace", "Fashion accessory", "Bicycle", "Calculator", "Camera", "Food processor", "Jug", "Mixing bowl", "Nightstand", "Oboe", "Oven", "Paper cutter", "Pencil case", "Perfume", "Pillow", "Personal care", "Pizza cutter", "Pressure cooker", "Printer", "Refridgerator", "High heels", "Skateboard", "Slow cooker", "Teddy bear", "Teapot", "Vase", "Wall clock", "Taco", "Tart", "Tea", "Waffle", "Wine", "Guacamole"]

Image Description: `[description]`

Answer:

Figure 14: **Prompt Templates for COMPLETENESS.** A prompt template for evaluating COMPLETENESS.

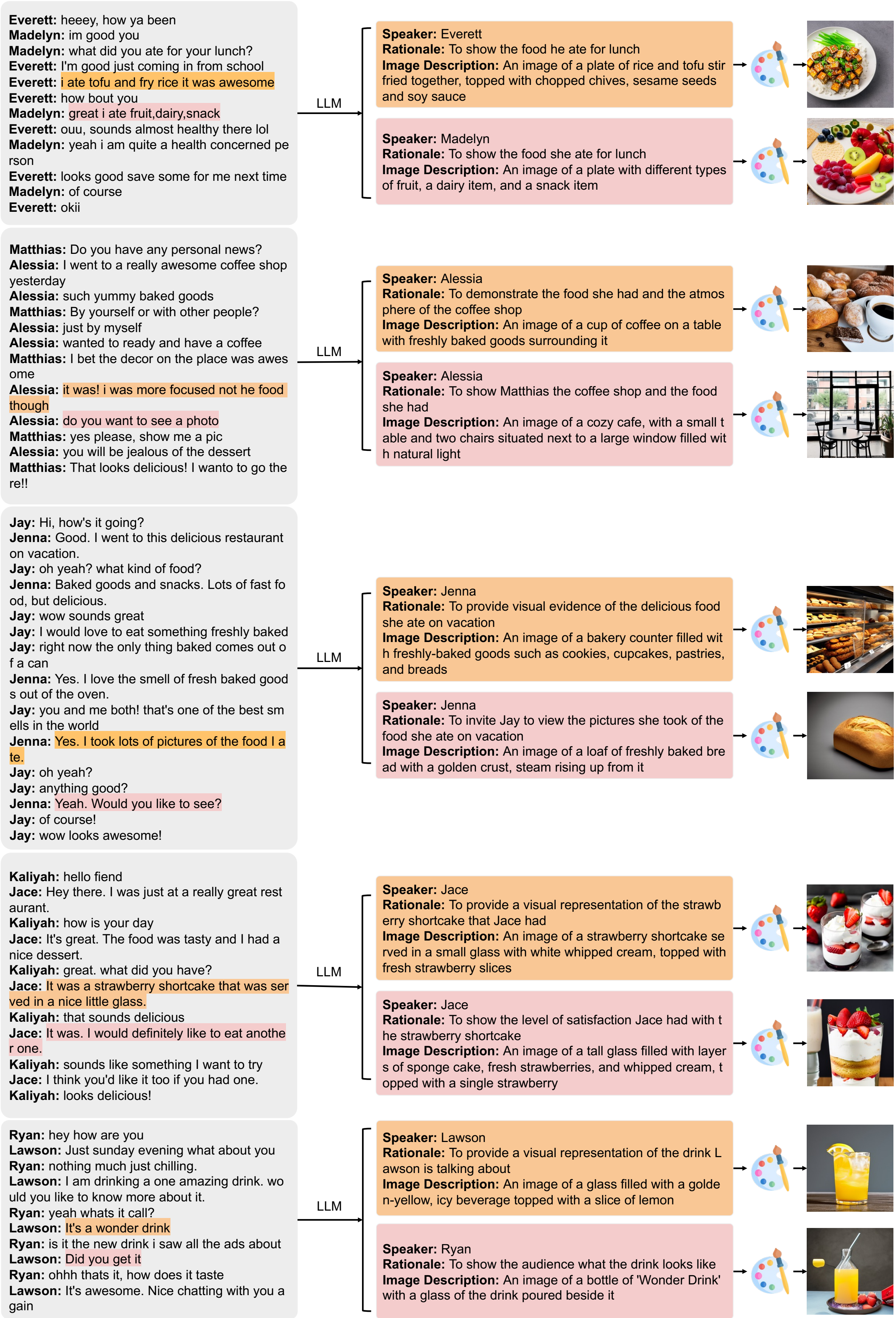

Figure 15: **More Examples of *aug*-PhotoChat .** We present more generated examples of *aug*-PhotoChat dataset using our proposed framework with LLM (i.e., `ChatGPT 1106`) and Stable Diffusion( ).