A BROADER IMPACTS

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 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.

B LIMITATIONS

Here, we highlight some limitations of our work. Firstly, our restriction-based 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.

C DISCUSSIONS

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 7'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.

Image Retrieval for Real Use Cases. We believe that utilizing a text-to-image retrieval model (e.g., CLIP (Radford et al., 2021)) to build an image-suggesting system would be more effective in actual use cases. The image generation step mentioned in our paper is used for creating the PhotoChat++ dataset, an application of our proposed framework, by generating images using StableDiffusion. Therefore, at a practical level, within our framework's stage 2, the LLM-generated image description can be used for text-to-image retrieval, allowing for the sharing of more diverse and up-to-date images from phone albums or the internet.

D Motivation behind Providing Full Context in Stage 1

In Stage 1, the full history is given as input. In Stage 2, only the previous conversation context is provided based on the image-sharing turn identified in Stage 1 (as depicted on the left side of Figure 4). Due to differing objectives, we allowed the LLM to view the entire future dialogue in Stage 1. This paper argues that multiple turns can serve as image-sharing turns (as detailed in Section 2). Therefore, we needed to identify other possible turns in the original PhotoChat dataset, which only had a single image-sharing turn. Given that the dialogue context was already fixed, it was crucial to find image-sharing turns that wouldn't disrupt the existing conversation flow after inserting relevant images at the image-sharing turns. Suppose we provided input to the LLM at every turn, similarly at Stage 2. In that case, there might be inconsistencies between the dialogue coherence after the image sharing and the subsequent dialogue content (which existed in the original PhotoChat). Thus, while giving the LLM the future dialogue as input can act as a hint, it was an inevitable choice in our experiments to validate our claim.

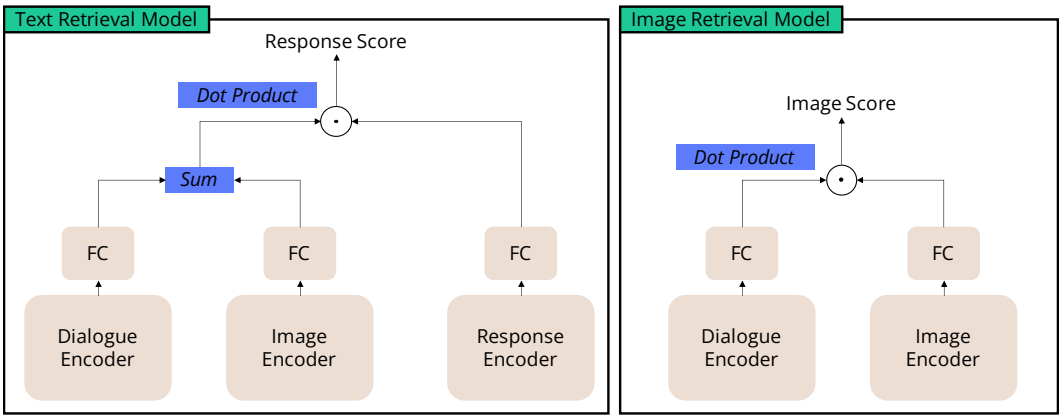

Figure 8: Architectures of two baseline models: Text retrieval and Image retrieval.

E Implementation Details of LLMs

To evaluate the *image-sharing* capabilities of LLMs, we call InstructGPT (Ouyang et al., 2022), 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.

F 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.

F.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.

F.2 Baseline Models

As illustrated in Figure 8, 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 comprises three main components: the dialogue encoder, the response encoder, and the image encoder. The dialogue encoder processes the entire dialogue history into a fixed-size representation. To achieve this, we use the BERT model (Devlin

et al., 2018). The dialogue history consists of up to three turns preceding the current turn. Each turn is concatenated using the `[SEP]` special token. The response encoder is responsible for converting the response into a fixed-size representation. While it also utilizes the BERT model, the specific BERT version used here differs from that employed in the dialogue encoder. After processing the text with BERT, we apply mean pooling to the text representations for both the dialogue and response encoders. The pooled representations are passed through a linear projection layer, followed by the ReLU activation function (Nair & Hinton, 2010). The image encoder is to extract feature vectors from images, and for this purpose, we utilize the CLIP-base model (Radford et al., 2021). Once the feature vectors are extracted from the dialogue and images, we perform an element-wise addition of the image vectors and dialogue vectors. We calculate the dot product between the response feature vector and the resulting summed vector to compute the loss.

Image Retrieval Model. The image retrieval model comprises two main components: the dialogue encoder and the image encoder. The dialogue encoder utilizes the BERT-base model to transform the dialogue into a representation. After encoding, we apply mean pooling to the text representations derived from this dialogue encoder. For image representation, we employ the CLIP-base model. Following the encoding processes, the image and dialogue vectors are passed through separate linear projection layers, each followed by a ReLU activation function. We calculate the dot product between the image feature vector and the dialogue vector to determine the loss.

F.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.

G Human Evaluation Questionnaire

We present a list of questions and multiple-choice options used for the human evaluation.

- Image-Sharing Turn Relevance: Do you think the image-sharing turn in the given dialogue is appropriate?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- Image-Sharing Speaker Adequacy: Do you think the speaker who shared the image in the given dialogue is appropriate? (When you selected 3 or 4 in Image-Sharing Turn Relevance)

  Options: 1: No / 2: Yes

- Image-Sharing Rationale Relevance: Do you think the reason for sharing the image in the given dialogue is valid? (When you selected 3 or 4 in Image-Sharing Turn Relevance)

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- Image Description Relevance: Given the dialogue context, how relevant do you think the image description is? (When you selected 3 or 4 in Image-Sharing Turn Relevance)

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

H HUMAN EVALUATION SYSTEM

As shown in Figure 9, we present a screenshot of human evaluation system. We implement the human evaluation system using Label Studio (Tkachenko et al., 2020-2022).

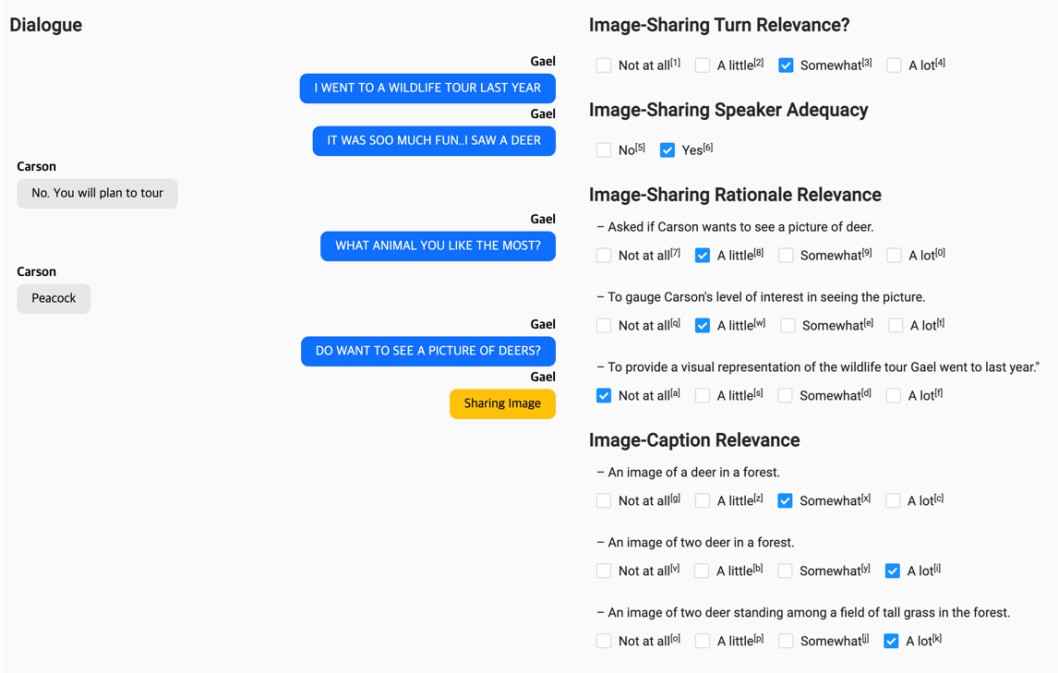

Figure 9: **Screenshot of the human evaluation system.**

I RATIONALE DISTRIBUTION

We present the rationale distribution as shown in Table 5.

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 5: **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.

J 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.

J.1 RESTRICTION-BASED PROMPT TEMPLATES

We present restriction-based prompt templates for each stage in our proposed framework, as shown in Figure 10.

J.2 PROMPT TEMPLATES FOR ABLATION STUDIES

We present prompt templates used in our ablation studies, as shown in Figure 11, Figure 12, Figure 13, and Figure 14.

K MORE EXAMPLES OF PHOTOCHAT++

We provide more examples of PhotoChat++ dataset.

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. Please list the selected utterances in descending order according to the confidence (0∼1) of your choice of how appropriate the utterance is for sharing an image now.

Dialogue:
`[dialogue]`

Restrictions:
(1) your answer should be in the format of "<UTTERANCE> | <CONFIDENCE> | <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 10: **Prompt Templates for Image-Sharing Behavior.** A prompt template for stage 1 (**top**). A prompt template for stage 2 (**bottom**).

Prompt Template for Stage 1 (w/o Restriction):
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. Please list the selected utterances in descending order according to the confidence (0∼1) of your choice of how appropriate the utterance is for sharing an image now. You should also provide a rationale for your decision, starting with "To". Your answer should be in the format of "<UTTERANCE> | <CONFIDENCE> | <SPEAKER> | <RATIONALE>". You MUST select the utterance in the given dialogue, NOT generate a new utterance.

Dialogue:
`[dialogue]`

Answer:
1.

Figure 11: **Prompt Templates for Stage 1 (w/o Restriction).** A prompt template used in the ablation study in Stage 1 for validating the effect of restriction.

Prompt Template for Stage 2 (w/o Restriction):
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. Your answer should be written starting with "An image of" and in one sentence. You do NOT include the speakerś name (i.e., `[speaker1]`, `[speaker2]`) in the image description. You should share a realistic image, NOT memes.

Dialogue:
`[dialogue]`

Image Description:

Figure 12: **Prompt Templates for Stage 2 (w/o Restriction).** A prompt template used in the ablation study in Stage 2 for validating the effect of restriction.

Prompt Template for One-Stage:

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 and describe the relevant image. Please list the selected utterances in descending order according to the confidence (0∼1) of your choice of how appropriate the utterance is for sharing an image now.

Dialogue:
`[dialogue]`

Restrictions:
(1) your answer should be in the format of "<UTTERANCE> | <CONFIDENCE> | <SPEAKER> | <RATIONALE> | <IMAGE DESCRIPTION>".
(2) you MUST select the utterance in the given dialogue, NOT generate a new utterance.
(3) the rationale should be written starting with "To".
(4) your answer should be written starting with "An image of" and in one sentence.
(5) you do NOT include the speaker's name (i.e., `[speaker1]`, `[speaker2]`) in the image description.
(6) you should share a realistic image, NOT memes.

Answer:
1.

Figure 13: **Prompt Templates for One Stage.** A prompt template for one stage framework.

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.

Figure 14: **Prompt Templates for Stage 1 (w/o Confidence-based Ranking).** A prompt template used in the ablation study in Stage 1 for validating the effect of confidence-based ranking.

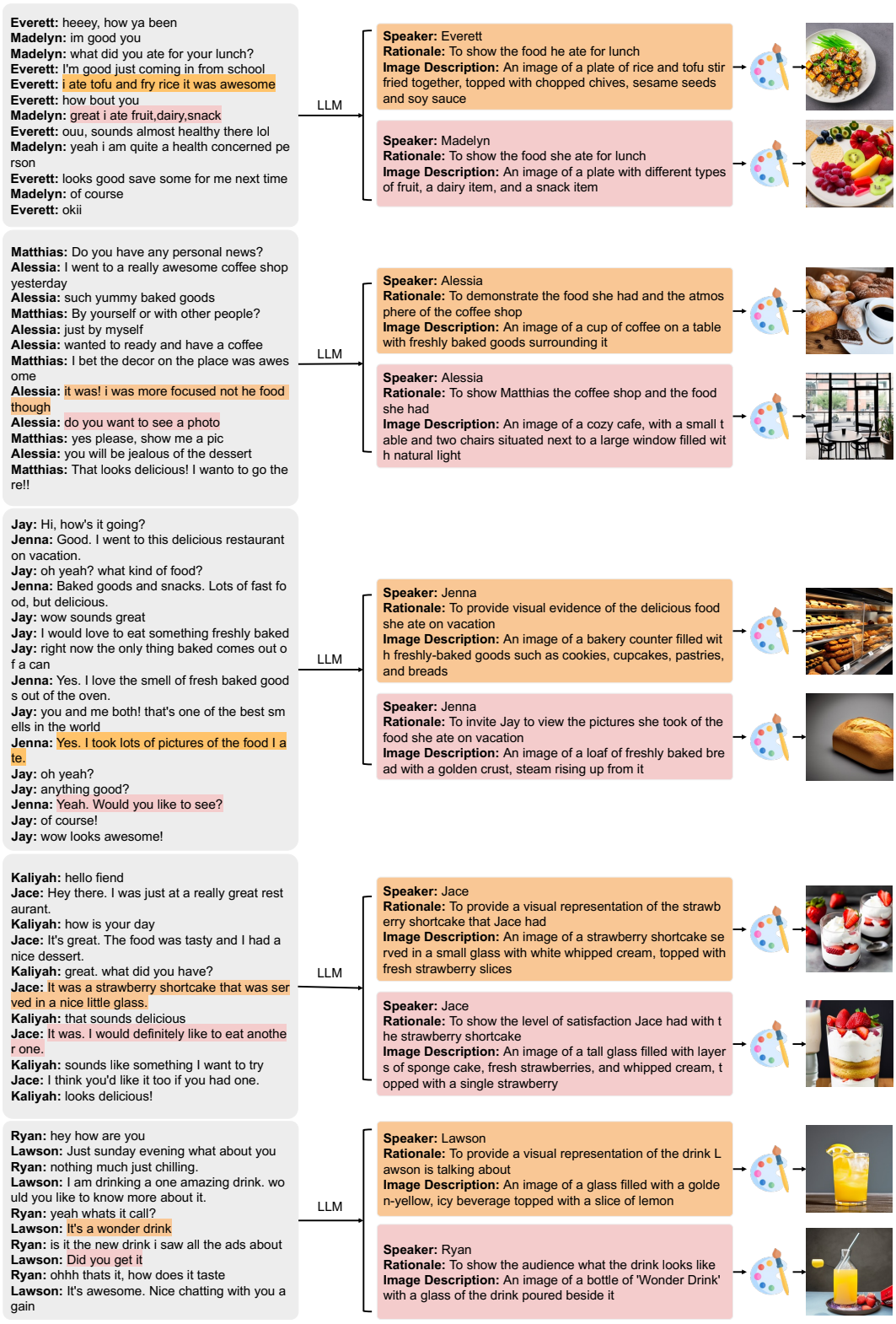

Figure 15: **More Examples of PhotoChat++.** We present more generated examples of PhotoChat++ dataset using our proposed framework with LLM (i.e., `text-davinci-003`) and Stable Diffusion(🦄).

L REGEX PATTERN

For stage 1, we use the regex pattern to extract the `utterance`, `confidence`, `speaker`, and `rationale` as follows:

```
^(?:\d+\.\s+)?\"?(?P<utterance>.*?)\"?\s+\|\s+(?P<confidence>.*?)\s+\|
\s+(?P<speaker>.*?)(?:\s+\|\s+(?:Rationale:\s+)?(?P<rationale>.*?))?$
```