Visualizing Dialogues: Enhancing Image Selection through Dialogue Understanding with Large Language Models

Chang-Sheng Kao Yun-Nung Chen
National Taiwan University, Taipei, Taiwan
r11922a14@csie.ntu.edu.tw y.v.chen@ieee.org

Abstract

Recent advancements in dialogue systems have highlighted the significance of integrating multimodal responses, which enable conveying ideas through diverse modalities rather than solely relying on text-based interactions. This enrichment not only improves overall communicative efficacy but also enhances the quality of conversational experiences. However, existing methods for dialogue-to-image retrieval face limitations due to the constraints of pre-trained vision language models (VLMs) in comprehending complex dialogues accurately. To address this, we present a novel approach leveraging the robust reasoning capabilities of large language models (LLMs) to generate precise dialogue-associated visual descriptors, facilitating seamless connection with images. Extensive experiments conducted on benchmark data validate the effectiveness of our proposed approach in deriving concise and accurate visual descriptors, leading to significant enhancements in dialogue-to-image retrieval performance. Furthermore, our findings demonstrate the method's generalizability across diverse visual cues, various LLMs, and different datasets, underscoring its practicality and potential impact in real-world applications.[1]

1 Introduction

In recent years, the landscape of online conversations has undergone a significant transformation thanks to the proliferation of instant messaging tools. Unlike the past, when these exchanges were confined to text alone, today's conversations have evolved into a multimodal experience, incorporating elements like images and speech. The various communication modes not only enhance engagement but also prove invaluable for conveying complex information that can be challenging to communicate solely through text. Sun et al. (2022) highlighted the advantages of integrating images into conversations. For example, when discussing a topic with someone who may not grasp the concept, sharing an image can provide visual clarity for better comprehension. Additionally, when precision is required to convey specific details about a subject, relevant images can be a more effective means of communication than text alone (Hsu et al., 2023). Consequently, the ability to generate responses using images is a crucial area of research in enhancing automatic dialogue systems. To equip these systems with the capacity to respond using images, a common method involves text-to-image retrieval, as demonstrated by previous work (Liao et al., 2018; Zang et al., 2021). In this approach, a model selects an appropriate image from a pre-constructed image repository based on the context of the ongoing conversation.

As storage costs decline and computational power advances, vision foundation models pre-trained on large-scale, open-domain image-text pairs have emerged (Radford et al., 2021; Jia et al., 2021; Yuan et al., 2021). These models have demonstrated outstanding performance in text-to-image retrieval tasks, excelling in both zero-shot and fully-trained scenarios. However, despite their impressive capabilities, these pre-trained vision-language models (VLMs) still come with some limitations. One significant limitation is their sub-optimal design for handling complete dialogue contexts effectively. Often, they suffer from extracting key information comprehensively from the entire conversation. Table 1 presents an illustrative example, where a dialogue-to-image model fine-tuned from CLIP (Radford et al., 2021) fails to correctly interpret the dialogue's intent. This highlights the challenge of dialogue comprehension, a task for which pre-trained VLMs may not be adequately equipped. Additionally, most existing VLMs typically impose input text length constraints during their pre-training stages, preventing them from processing the entirety of the dialogue context directly.

[1] https://github.com/MiuLab/VisualDialog

Findings of the Association for Computational Linguistics: ACL 2024, pages 11777–11788
August 11-16, 2024 ©2024 Association for Computational Linguistics

Dialogue context
B: how are you doing?
A: I'm doing good. Just out at a restaurant taking pictures for customers.
B: congratulations
A: It's hilarious watching people try to use chopsticks
B: i'm really happy for you friend
B: yeah, its really funny
A: Yeah, it's better than most gigs I get
B: even i still try to try to find a way around that thing
A: I give up and ask for a fork. I want that rice in my mouth!!!!!
A: (share a photo)

Ground-truth

Retrieved top-1

Dialogue-associated visual cues

- main subject: customers
- foreground objects: chopsticks, table, food
- background scene: restaurant
- events: eating food

Table 1: An example of a dialogue and the shared image; the fine-tuned CLIP model fails to retrieve the correct one. Red indicates the missing elements, blue indicates a perfect match, and orange suggests a partial match.

This constraint can lead to the loss of crucial contextual information, potentially undermining the model's overall performance.

Inspired by Menon and Vondrick (2022), we leverage the reasoning capabilities of large language models (LLMs) to generate the visual descriptor for the dialogue context. These descriptors encapsulate speculations about the image that the speaker intends to share, aiming to provide concise and precise cues for better text-to-image retrieval. Our objective is to address the aforementioned limitations and enhance task performance. Given that most vision models excel at identifying objects, scenes, and other visual elements in images (Kuznetsova et al., 2020; Zang et al., 2021), we employ a set of visually-focused queries, such as *main subject* and *background scene*, to bridge the gap between the ongoing dialogue and the pool of potential image candidates. These queries serve as templates for the LLM to predict corresponding visual cues based on the dialogue context. We then utilize these queries and their resulting answers as dialogue-associated visual descriptors, as illustrated in the bottom part of Table 1. Our experiments on the benchmark dataset showcase the exceptional performance of our approaches, surpassing all previous results. In addition to demonstrating the effectiveness of our LLM-generated visual descriptor, we compare it with other descriptor creation methods and conduct an in-depth analysis to evaluate the efficacy of each proposed query.

Our contributions can be summarized as 3-fold:

- This paper introduces a novel approach for retrieving associated photos in dialogue systems, leveraging the reasoning capabilities of LLMs to generate visually-focused cues for improved image retrieval.
- We design a series of visually-focused queries based on common image features, employing them to construct conversation descriptors. Our experiments validate the effectiveness of these designed queries.
- The proposed approach achieves state-of-the-art performance on multiple benchmark datasets.

2 Related Work

Multimodal Dialogue Systems Recent years have witnessed a notable shift in research towards multimodal dialogues, moving beyond the confines of text-only interactions (Liu et al., 2022). While the exploration of image-grounded conversations, where textual dialogues are generated from images, has gained traction (Yang et al., 2021; Shuster et al., 2021; Chen et al., 2023; Gong et al., 2023; Zhu et al., 2023; Liu et al., 2023), an increasing number of studies are delving into the incorporation of multimodal responses within dialogue systems. This multimodal evolution enables human-machine conversations to reflect real-life human-human interactions and communicate concepts that are difficult to convey through text alone. For instance, Liao et al. (2018) introduced a task-oriented multimodal dialogue system featuring a taxonomy-based learning module that captures nuanced visual semantics and employs reinforcement learning to ensure response coherence. Further advancements in robotic technology now enable robots to more effectively understand and respond to human requests in life support scenarios. This requires the interpretation of both the visual context of the environment and the user's verbal communication, integrating multimodal capabilities into dialogue scenarios (Tanaka et al., 2024; Tsai et al., 2024). Moreover, Sun et al. (2022) introduced a frame-

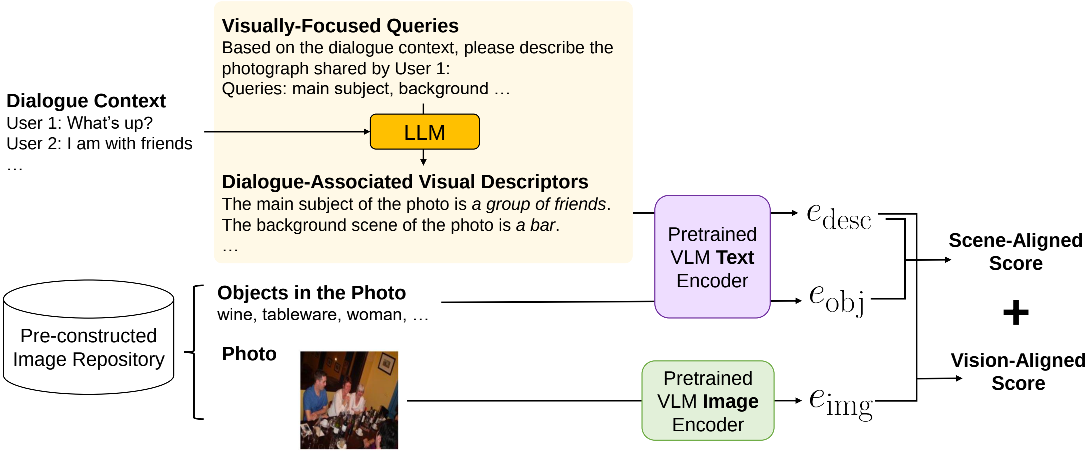

Visually-Focused Queries
Based on the dialogue context, please describe the photograph shared by User 1:
Queries: main subject, background …
Dialogue Context
User 1: What's up?
User 2: I am with friends
…
LLM
Dialogue-Associated Visual Descriptors
The main subject of the photo is *a group of friends*.
The background scene of the photo is *a bar*.
…
Pretrained VLM Text Encoder
e_{desc}
e_{obj}
Scene-Aligned Score
+
Vision-Aligned Score
Pre-constructed Image Repository
Objects in the Photo
wine, tableware, woman, …
Photo
Pretrained VLM Image Encoder
e_{img}

Figure 1: The framework of our proposed method. We employ the text encoder from a pre-trained VLM to encode both the descriptor and the object list. This yields two distinctive features, namely the descriptor embedding (e_{desc}) and the object list feature (e_{obj}). Additionally, we utilize the VLM's image encoder to process and encode the image, resulting in the image embedding (e_{img}). The final retrieval score is then computed by aggregating a scene-aligned score and a vision-aligned score.

work capable of directly generating multimodal responses via a text-to-target-modality generator. In contrast, rather than directly generating multimodal responses, Zang et al. (2021) achieve multimodal responses by employing image retrieval models to select appropriate images from a pre-existing image repository. For better practicality, our paper centers on the same task—recommending a suitable image from the user's image repository based on the ongoing dialogue context.

External Knowledge of LLMs for Visual Tasks Many studies have showcased that the commonsense knowledge and reasoning capabilities in large language models (LLMs) can significantly augment the performance of visual tasks. For instance, Tsimpoukelli et al. (2021) confirmed that by projecting image encodings into the embedding space of an LLM, it becomes possible to harness the rich knowledge contained within the LLM for few-shot visual question answering (VQA) tasks. Similarly, Zeng et al. (2022) introduced Socratic Models, which leverages multiple pre-trained large models trained on data from diverse domains. By translating non-language domain information into textual prompts, Socratic Models achieve state-of-the-art results in zero-shot image captioning and video-to-text retrieval tasks. Furthermore, Menon and Vondrick (2022) took a novel approach by obtaining visual features for different categories through queries to GPT-3 (Brown et al., 2020) using category names. These textual descriptors are then employed as internal representations for zero-shot visual classification and text-to-image retrieval tasks. Our work centers on harnessing the reasoning capabilities of LLMs to derive contextually relevant visual descriptions for shared photos within the dialogue context. Different from the prior work based on non-language domains or single sentences, our approach focuses on the nuanced domain of photo sharing within conversations, which presents unique challenges due to its reliance on commonsense knowledge and an understanding of human-human interactions.

3 Methodology

Our objective is to select an image from a pre-constructed photo set $\{(v_j, o_j)\}_{j=1}^{m}$ given a dialogue context D, where v_j represents an image candidate and o_j lists the objects appearing in v_j. Note that the object lists can be obtained through object detection in the pre-processing stage, and we treat this object information as given data.

Figure 1 illustrates the proposed framework, which introduces an innovative approach to estimate retrieval scores for each image candidate within a dialogue context. These scores are based on two criteria: **scene-aligned** and **vision-aligned** scores, both relying on visual descriptors. The scene-aligned score assesses whether the speculated visual cues align with the image-associated objects in a *textual* format. In contrast, the vision-aligned score evaluates the alignment between the visual description and the image using *vision-language* models.

Visual Features & Descriptions	**Examples**
Main subject: the photo-focused objects for conveying a particular theme	*people, cakes, buildings*
Prominent objects in the foreground: objects in addition to the main subject convey signal for photo understanding.	*a bar counter and bottles in a photo taken at a bar*
Background scene: the background scene in the photo	*restaurants, bars, outdoors*
Events: activities or events currently captured in the photo	*weddings, birthdays, eating food*
Materials and attributes: finer details about the photo	*teapot made of ceramic, black and white feathers*

Table 2: All designed descriptors used in the proposed method.

3.1 Dialogue-Associated Visual Descriptor

Considering that visual descriptors can significantly enhance the understanding of visual content (Menon and Vondrick, 2022), we focus on generating dialogue-associated visual descriptors to improve image retrieval capabilities. To create high-quality visual descriptors that can connect with visual elements in the photo, we define a set of visually-focused queries, denoted as $Q = \{q_i\}$. These queries encompass various visual attributes related to an image, such as *main subject* and *background scene*, which are instrumental in linking the target photo to the dialogue.

Drawing from prior work (Kuznetsova et al., 2020; Zang et al., 2021) and our common experiences, we assume that photos shared in online messaging typically contain components such as *main subjects*, *prominent foreground objects*, *background scenes*, *events*, and *materials and attributes*, as detailed in Table 2. Note that we do not expect all answers to these queries to be perfectly extracted from the dialogue context or found in the ground-truth image. Instead, our goal is to leverage automatically inferred visual descriptors to bridge the gap between the image and the given dialogue context.

Leveraging the powerful reasoning capabilities of large language models (LLMs) (Touvron et al., 2023), we construct a prompt comprising the dialogue D and the set of queries Q and input it into the LLM. This process yields a set of dialogue-associated visual descriptors in a zero-shot manner:

$$\text{desc} = \text{LLM}(D, Q). \tag{1}$$

For instance, a generated visual descriptor regarding the main subject might read, "*The main subject of the photo is a group of friends*." The used prompts can be found in Appendix A.

3.2 Image Relevance Estimation

To measure the relevance of each image candidate in the context of a given dialogue D, we calculate two retrieval scores based on their generated visual descriptors desc: $S_{\text{scene}}(o_j, \text{desc})$ and $S_{\text{vision}}(v_j, \text{desc})$. The former score assesses if the objects in the photo candidate align with the inferred visual descriptors in their *text-only* forms, referred to as the **scene-aligned** score. The latter score evaluates if the photo candidate matches the visual descriptions through *multimodal* methods, termed the **vision-aligned** score.

3.3 Image Retrieval Learning

Our task involves retrieving the target image from a pre-constructed photo set, and it can be approached in two settings: 1) zero-shot and 2) training with contrastive leanrning.

3.3.1 Zero-Shot

Using the descriptor desc derived from the dialogue context D, we employ a pre-trained vision language model (VLM) for zero-shot text-to-image retrieval. This process yields two scores through its text encoder and image encoder, as illustrated in Figure 1. The final retrieval score is calculated as:

$$S_{\text{scene}}(o_j, \text{desc}) + \lambda \cdot S_{\text{vision}}(v_j, \text{desc}), \tag{2}$$

where λ is a weighting parameter. The image with the highest score is selected in a zero-shot manner.

3.3.2 Contrastive Learning

To further enhance retrieval performance, we finetune the VLM model using the training set. Following the pre-training stage outlined by Radford et al. (2021), we apply contrastive learning to optimize our dialogue-image retriever. During training, we randomly sample a minibatch of dialogue-associated descriptors and photo pairs, designating (desc, v^*, o^*) as the positive example, while the remaining $(b - 1)$ examples within the minibatch serve as negative examples. The contrastive losses are calculated separately for the scene and vision components, focusing on aligning dialogue-associated visual descriptors and the target photo.

$$\mathcal{L}_{\text{scene}} = -\log \frac{\exp(S_{\text{scene}}(o^*, \text{desc})/\tau)}{\sum_{j\in b} \exp(S_{\text{scene}}(o_j, \text{desc})/\tau)},$$

$$\mathcal{L}_{\text{vision}} = -\log \frac{\exp(S_{\text{scene}}(v^*, \text{desc})/\tau)}{\sum_{j\in b} \exp(S_{\text{scene}}(v_j, \text{desc})/\tau)},$$

where τ is the trainable temperature parameter. The final training loss is a combination of these contrastive losses:

$$\mathcal{L} = \frac{1}{b} \sum_{j\in b} (\mathcal{L}_{\text{scene}} + \lambda \cdot \mathcal{L}_{\text{vision}}), \tag{3}$$

where λ is a weighting parameter. This approach optimizes our dialogue-image retrieval model through contrastive learning.

4 Experiments

For the LLM in (1), we utilized well-established LLMs with instruction tuning and reinforcement learning from human feedback (RLHF), including LLaMA-2-7B-Chat and LLaMA-2-13B-Chat (Touvron et al., 2023). We employed greedy decoding for generating descriptors to ensure the correct format and reasoning capability. Our pre-trained vision-language model (VLM) backbone is CLIP ViT-B/32, and VLM training is executed on a single NVIDIA GeForce RTX 2080 Ti GPU with a batch size of 56. We utilize the ADAM optimizer with an initial learning rate of 1e-5. The weighting parameter λ was set to 1 to strike a balance between scene-alignment and vision-alignment.

4.1 Data & Metrics

Our approach is evaluated on a benchmark multimodal dialogue dataset called PhotoChat. PhotoChat dataset (Zang et al., 2021) is characterized by open-domain, high-quality multimodal dialogues and comprises 10,917 images paired with 12,286 dialogues. Specifically, the dataset is divided into 10,286 instances for training, 1,000 for validation, and another 1,000 for testing. Each image in the dataset is accompanied by an associated object list presented in textual form. In each data instance, one photo is shared within the context of the conversation.

Given that this task can be formulated as an image retrieval task, we employed $Recall@k$ ($R@k$) as our evaluation metric. During the training phase, we select the final model based on the highest $avg(R@1, R@5, R@10)$ score on the validation set. In the testing phase, for each dialogue instance, the trained models retrieved images from the candidate photos in the testing set.

4.2 Baselines

We compare our approach against several established baselines:

- VSE++: Faghri et al. (2018) incorporated hard negatives in the ranking loss to learn visual-semantic embeddings for text-image retrieval.
- SCAN: Lee et al. (2018) utilized stacked cross attention to align image regions and words in a sentence and calculate image-text similarity.
- Dual Encoder (DE): Previous work (Parekh et al., 2021; Zang et al., 2021) employed a dual encoder architecture, where one encoder processes the image and its object list using CLIP ViT-B/32 for images and FFNN for object features. For the dialogue encoder, two different text encoders were experimented with: CLIP ViT-B/32 Text and BERT (Devlin et al., 2019) with an additional projection to ensure consistent dimensions. The retrieval similarity between the image and dialogue encodings is measured using dot product.

4.3 Descriptor Variants

In addition to the query-based descriptors, we conduct experiments using the following descriptor variants for in-depth analysis:

- **Desc - Diag** (whole dialogue as descriptors): All utterances are concatenated to form the descriptors, allowing the image retriever to utilize complete cues within the dialogue.
- **Desc - Caption** (caption as descriptors): Inspired by Li et al. (2023), we performed zero-shot image captioning on images in the training set using BLIP-2. We then trained a text generator to create image captions as descriptors based on a given dialogue.
- **Desc - Summary** (summary as descriptors): Descriptors are generated by LLMs based on a dialogue summary, offering a more concise representation of the conversation.
- **Desc - Guessing** (visually-focused guessing as descriptors): LLMs are allowed to speculate about the features of the upcoming shared photo from the dialogue without being constrained by a specific query.
- **Desc - Queries** (visually-focused query descriptors): Utilizing our designed visually-focused attributes as descriptors.

Method	LLM	Zero-Shot				Fully-Trained			
		R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg
VSE++†	-	-	-	-	-	10.20	25.40	34.20	23.27
SCAN†	-	-	-	-	-	10.40	27.00	37.10	24.83
DE - Diag (BERT)	-	-	-	-	-	12.88	35.13	47.75	31.92
DE - Diag (CLIP)	-	-	-	-	-	14.76	35.78	47.12	32.55
Desc - Diag	-	16.00	30.90	37.70	28.20	40.35	58.77	66.88	55.33
Desc - Caption	BLIP-2	-	-	-	-	16.68	35.34	45.17	32.40
Desc - Summary	LLaMA-2-7B-Chat	22.90	40.10	47.60	36.87	42.81	62.42	71.35	58.86
Desc - Summary	LLaMA-2-13B-Chat	24.40	40.50	48.30	37.73	**44.17**	64.23	72.66	60.35
Desc - Guessing	LLaMA-2-7B-Chat	27.60	47.80	58.10	44.50	42.55	64.22	72.29	59.69
Desc - Guessing	LLaMA-2-13B-Chat	**29.30**	**51.30**	**59.80**	**46.80**	43.18	**65.45**	73.43	60.69
Desc - Queries	LLaMA-2-7B-Chat	22.60	42.20	50.40	38.40	37.34	57.52	66.62	53.83
Desc - Queries	LLaMA-2-13B-Chat	26.40	45.80	55.10	42.43	44.00	64.78	**73.95**	**60.91**

Table 3: Retrieval performance for zero-shot and fully-trained settings (%). We employ the LLM with greedy decoding to ensure the correct format and reasoning capability. Each number is the average over 10 runs with different random seeds. †denotes that we directly report the numbers from Zang et al. (2021).

4.4 Results

Table 3 provides a comprehensive overview of the results for both zero-shot and fully-trained settings on PhotoChat dataset. In zero-shot scenarios, **Desc - Guessing** emerges as the top-performing method among all results. Notably, **Desc - Queries** outperforms **Desc - Summary**, indicating that visually-focused queries and guessing contribute valuable information for linking the desired images. We hypothesize that the poor performance of **Desc - Summary** is because the LLM summarize all content of the dialogue, rather than focusing solely on the visual information. Consequently, **Desc - Summary** often includes sentences that do not aid in text-to-image retrieval. As for why **Desc - Queries** did not perform better than **Desc - Guessing**, we hypothesize that it is because we construct our descriptors using the templates from 3.1. These sentences tend to be less fluent and are not the types of sentences CLIP was familiar with during pre-training stage. In contrast, **Desc - Guessing** generates more fluent descriptors without any constraints, making them closer to the sentences seen by CLIP during pre-training stage. This gives **Desc - Guessing** an advantage in zero-shot scenarios.

In the fully-trained setting, the descriptor-based results (**Desc - Summary**, **Desc - Guessing**, **Desc - Queries**) with LLaMA-2-13B-Chat exhibit similar performance, with **Desc - Queries** achieving the highest average performance. These results validate the effectiveness of our proposed approach, demonstrating that the generated visual descriptors successfully facilitate the connection between associated images through the LLM's understanding of dialogue. Additionally, it is evident that LLaMA-2-13B-Chat outperforms LLaMA-2-7B-Chat due to its stronger reasoning abilities for understanding dialogues. When compared to the fully-trained baselines, our proposed descriptor-based methods achieve superior performance even in zero-shot settings, establishing a new state-of-the-art performance achieved by a single model.

Ensemble	R@1	R@5	R@10	Avg
S + G	47.32	69.62	77.63	64.86
S + Q	47.78	68.81	77.61	64.73
G + Q	47.44	68.90	77.15	64.50
S + G + Q	48.79	70.01	78.44	65.75
S + G + Q + C	**48.84**	**70.20**	**78.74**	**65.93**

Table 4: Ensemble results of fully-trained retrievers with LLaMA-2-13B-Chat as the LLM (%). (S: Summary; G: Guessing; Q: Queries; C: Caption).

4.5 Ensemble

We further conduct experiments on ensemble learning using all descriptor-based results based on the validation set. The results in Table 4 demonstrate that ensemble learning consistently improves performance. Even in cases where the caption model performs poorly in a fully-trained setting, ensemble learning benefits other models. These findings highlight the efficacy of combining various types of descriptors, leading to the best overall performance and establishing a new state-of-the-art for PhotoChat. This suggests that the generated descriptors focus on diverse patterns that can complement each other and enhance scores.

Method	**Score**	**R@1**	**R@5**	**R@10**	**Avg**
Desc - Summary	Scene-Aligned (Text-Only)	35.07	49.37	57.66	47.37
	Vision-Aligned (Multimodal)	29.37	53.18	62.49	48.35
Desc - Guessing	Scene-Aligned (Text-Only)	**35.82**	50.58	58.30	48.23
	Vision-Aligned (Multimodal)	28.41	53.78	63.90	48.70
Desc - Queries	Scene-Aligned (Text-Only)	35.53	50.64	58.68	48.28
	Vision-Aligned (Multimodal)	29.16	**54.28**	**64.17**	**49.20**

Table 5: The results of the model trained using either scene-aligned or vision-aligned scores.

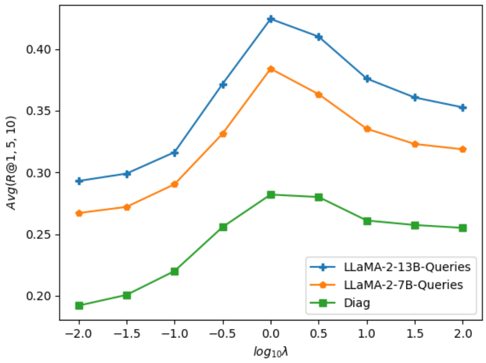


Figure 2: Results of different λ in zero-shot scenarios. A smaller λ indicates greater reliance on the scene-aligned score, while a larger λ indicates greater reliance on the vision-aligned score.

Error Rate	**Type**	**R@1**	**R@5**	**R@10**	**Avg**
0%	-	26.40	45.80	55.10	42.43
15%	missing	24.75	43.85	53.45	40.68
	incorrect	22.04	39.99	49.01	37.01
	both	23.29	42.13	50.74	38.72
25%	missing	23.05	42.17	51.50	38.91
	incorrect	19.93	*37.41*	*45.59*	*34.31*
	both	21.82	39.92	48.42	36.72
35%	missing	21.91	40.61	49.89	37.74
	incorrect	*17.24*	*33.87*	*41.66*	*30.92*
	both	*19.25*	*36.98*	*45.33*	*33.85*
100%	missing	19.80	38.50	46.60	34.97

Table 6: Sensitivity of diverse object error rates of LLaMA-2-13B-Chat in zero-shot settings. **Both** denotes the random removal or replacement of objects. 100% missing means that we use only the vision-aligned score for retrieval. The italicized fonts indicate scores lower than vision-aligned only results.

5 Analysis

5.1 Effectiveness of Two Alignment Scores

Our proposed method incorporates two scores: scene-aligned (text-only) and vision-aligned (multimodal) scores. We conduct an ablation study to assess the impact of each score. Table 5 presents the experimental results. The results show that models trained solely on the scene-aligned score (text-only) perform better in terms of $R@1$, whereas models trained on the vision-aligned score (multimodal) perform better for $R@5$ and $R@10$.

In addition, we conducted experiments with different weighting parameter λ in (2). Figure 2 illustrates the results obtained for various λ values in the zero-shot scenarios. It can be observed that the best performance is achieved when λ approaches 1. This indicates that excessive bias towards either the scene-aligned score or the vision-aligned score leads to a decrease in performance. Interestingly, our findings reveal that the optimal zero-shot performance occurs at $\lambda = 1.2$.

5.2 Object Detection Sensitivity

Given our experimental assumption of having available object lists for the images, a degradation in system performance may occur when object detection is poor. Consequently, we explored the sensitivity of error propagation in our proposed method by simulating various error rates in detected objects. Specifically, we randomly removed objects from the gold object list or substituted them with others from the training set to simulate scenarios of “missing” or “incorrect” objects. As shown in Table 6, it is evident that with increasing error rates, regardless of the error type, performance tends to degrade. This suggests that our method is still influenced by error propagation. Moreover, the impact of incorrect objects is particularly severe since they may mislead the model in image selection. Nevertheless, even with a 35% error rate, our results outperform relying solely on the vision-aligned score, indicating a relative robustness to missing objects. Therefore, precision in object detection during object list extraction is more important for our proposed approach.

5.3 Visually-Focused Query Impact

To assess the influence of different visually-focused queries on our results, we conducted experiments by systematically removing individual queries from the original query set (see Table 7). Remarkably, the query concerning the *main subject* emerges as the most pivotal feature for bridging the dialogue

Method	R@1	R@5	R@10	Avg
Original	44.00	64.78	**73.95**	60.91
- main subject	28.80	49.16	58.41	45.56
- foreground objects	40.44	61.62	70.49	57.52
- background scene	43.65	64.05	72.88	60.19
- events	42.91	64.00	72.78	59.90
- materials & attributes	43.22	64.60	73.59	60.47
+ atmosphere or mood	43.67	64.89	73.85	60.80
+ lighting	**44.13**	**64.95**	73.93	**61.00**

Table 7: The results of different queries. All additions and removals are based on the original query set.

Method	VisDial	MMDialog
Koh et al. (2023)	40.57	-
Desc - Diag	22.53	32.47
Desc - Summary	35.77	19.03
Desc - Guessing	**49.29**	36.93
Desc - Queries	44.45	36.40
Desc - Queries (all)	43.80	**38.27**

Table 8: Average zero-shot performance on VisDial and MMDialog (%).

context and the target image, as its removal leads to a significant decrease in scores. Following in descending order of impact are the queries related to *foreground objects*, *events*, *background scene*, and *materials and attributes*.

Additionally, we augmented the original query set with two common photo features: *atmosphere or mood* and *lighting*.

- **Atmosphere or mood:** Research by Sun et al. (2022) suggests that photographs convey not only object details but also emotions and ambiance, evoking feelings like happiness, boredom, and coziness.
- **Lighting:** As highlighted by Hunter et al. (2021), the presence and quality of light are fundamental in photographic composition. Our investigation explores whether an LLM can accurately predict lighting conditions within a photo solely based on dialogue context and whether this predictive information enhances text-image retrieval capabilities.

Results indicate that *atmosphere or mood* improves performance at $R@5$, while *lighting* enhances results at both $R@1$ and $R@5$ compared to the original set. This suggests that these abstract and challenging-to-predict queries have varying impacts on performance.

5.4 Generalization to Other Datasets

To explore the generalizability of our approach, we conducted experiments on other benchmark datasets, VisDial (Das et al., 2017) and MMDialog (Feng et al., 2023), with identical settings.

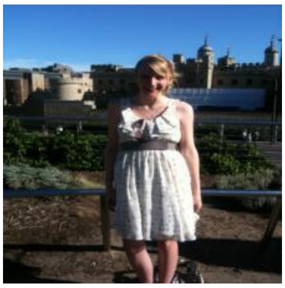
- VisDial (Das et al., 2017): Each instance in VisDial comprises a single image from the COCO dataset (Lin et al., 2014) and a dialogue context consisting of question-answer pairs. During testing, the model's task is to retrieve the correct image from a pool of 2,064 images in the testing set.
- MMDialog (Feng et al., 2023): This dataset comprises multimodal open-domain dialogues gathered from a global social media platform. Instances in the testing set are linked with 999 corresponding negative images. MMDialog uniquely allows dialogue context to contain both text and images. In our experiments, we employed BLIP2 (Li et al., 2023) for image captioning within the dialogue context. These resulting captions replaced the images, enabling textual dialogue context input into the LLM. Our experiments were conducted on a random sample of 1,000 instances from the testing set.

The average performance of LLaMA-2-13B-Chat in zero-shot scenarios across different datasets is summarized in Table 8, with detailed scores available in Appendix B. Our findings indicate that our proposed approach consistently outperforms **Desc - Diag** baseline and even recent work (Koh et al., 2023), showcasing its robust generalization capability across diverse datasets.

5.5 Qualitative Study

To further explore the quality of our generated descriptors, we conducted a qualitative analysis across various models and descriptors. We present an example of a testing dialogue using different LLMs and descriptor generation methods in Table 9. A comparison between LLaMA-2-7B-Chat and LLaMA-2-13B-Chat in the **Desc - Queries** domain revealed that LLaMA-2-7B-Chat often delivers inaccurate or unrelated answers (highlighted in red). In contrast, LLaMA-2-13B-Chat tends to produce more precise answers and demonstrates an ability to infer information not explicitly present in the dialogue context (e.g., medieval architecture). This discrepancy is likely attributable to differences in reasoning capabilities between the two models.

Both **Desc - Summary** and **Desc - Guessing** can accurately describe the features of the photos. However, **Desc - Summary** occasionally incorporates additional details not directly related to the

Dialogue Context	**Ground-Truth Photo**
B: whats up **A:** Hanging out with my student Maren **B:** Oh thats cool **B:** how was it ? **A:** Great. We are sightseeing castles **B:** That sounds really cool actually **B:** took any pictures of it? **A:** Yes **B:** Show me it I wanna see **A:** sure **A:** (share a photo)	

LLaMA-2-7B-Chat		
Desc - Queries	**Desc - Summary**	**Desc - Guessing**
The main subject of the photo is a picture of a castle. The prominent objects in the foreground of the photo is Maren and **A**. The background scene of the photo is a beautiful landscape with rolling hills and a blue sky. The events of the photo is sightseeing. The materials and attributes of the photo is a camera and a smartphone	**A** is a teacher who is spending time with their student Maren. They are sightseeing castles and took pictures of them. **A** is willing to show the pictures to **B**.	It is likely that **A** shared a photograph of a castle that they and Maren visited during their sightseeing trip. The photograph may show the exterior or interior of the castle, or a specific feature or detail of the castle that caught **A**'s attention.
LLaMA-2-13B-Chat		
Desc - Queries	**Desc - Summary**	**Desc - Guessing**
The main subject of the photo is Maren. The prominent objects in the foreground of the photo are castles. The background scene of the photo is countryside. The events of the photo are sightseeing. The materials and attributes of the photo are medieval architecture.	**A** is a teacher or mentor figure, as they are referred to as "my student Maren" **A** and Maren are currently sightseeing castles together **A** has taken pictures of the castles they have visited **A** is willing to share these pictures with **B**	**A** shared a photograph of a castle that they visited with their student Maren. The photograph is likely to be a scenic view of the castle, showcasing its architecture and surroundings. The photograph may include Maren in the frame, possibly posing in front of the castle or looking out at the view.

Table 9: Red text represents incorrect information. Orange text represents redundant sentences that do not aid in the text-to-image retrieval task.

photos, such as "*A is willing to share these pictures with B*," likely because we instructed the LLM to summarize information about the users sharing the photo rather than solely focusing on the photo-related information. Our analysis suggests that our visually-focused approaches are particularly effective for this task, especially when coupled with the powerful reasoning capabilities of LLMs.

6 Conclusion

This paper presents a novel approach to enable multimodal dialogue systems for seamless photo sharing. By harnessing the reasoning capabilities of LLMs, our method generates precise visual cues from ongoing dialogue contexts. Addressing challenges encountered in prior methods, such as accurately understanding extensive dialogue contexts and handling input length constraints, our approach exhibits clear superiority in experimental results. Additionally, our comprehensive ablation study confirms the effectiveness of text-only visual descriptors, indicating a promising pathway for bridging intricate dialogues and images through deep dialogue understanding via LLMs. This work not only advances photo sharing within dialogues but also lays groundwork for future sophisticated multimodal dialogue systems.

Limitations

Our method assumes the availability of object detection capabilities during pre-processing to extract object lists associated with the images. This reliance on object detectors may limit the method's applicability in scenarios where object detection is challenging or unavailable, potentially affecting its performance.

Lastly, our method assumes that the shared images align with the given dialogue context. In cases where users share images that are intentionally misleading, unrelated to the conversation, or dependent on users' personal information, our method may struggle to retrieve appropriate images, leading to potential accuracy issues in such scenarios.

Acknowledgements

We thank the reviewers for their insightful comments. This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 111-2222-E-002-013-MY3, 111-2628-E-002-016, and 112-2223-E002-012-MY5. We thank to National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Xiaolin Chen, Xuemeng Song, Liqiang Jing, Shuo Li, Linmei Hu, and Liqiang Nie. 2023. Multimodal dialog systems with dual knowledge-enhanced generative pretrained language model. *ACM Transactions on Information Systems*, 42(2):1–25.

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. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363, Toronto, Canada. Association for Computational Linguistics.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans.

Tsu-Yuan Hsu, Chen-An Li, Chao-Wei Huang, and Yun-Nung Chen. 2023. Visually-enhanced phrase understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5879–5888.

Fil Hunter, Steven Biver, Paul Fuqua, and Robin Reid. 2021. *Light—science & magic: An introduction to photographic lighting*. Routledge.

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.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pages 17283–17300. PMLR.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216.

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

Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 801–809.

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

Guangya Liu, Shiqi Wang, Jianxing Yu, and Jian Yin. 2022. A survey on multimodal dialogue systems: Recent advances and new frontiers. In *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, pages 845–853.

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

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

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2021. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2855–2870, Online. Association for Computational Linguistics.

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.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2021. Multi-modal open-domain dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4863–4883, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2022. Multimodal dialogue response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866, Dublin, Ireland. Association for Computational Linguistics.

Shohei Tanaka, Konosuke Yamasaki, Akishige Yuguchi, Seiya Kawano, Satoshi Nakamura, and Koichiro Yoshino. 2024. Do as i demand, not as i say: A dataset for developing a reflective life-support robot. *IEEE Access*.

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*.

Shang-Chi Tsai, Seiya Kawano, Angel F Garcia Contreras, Koichiro Yoshino, and Yun-Nung Chen. 2024. ASMR: Augmenting life scenario using large generative models for robotic action reflection. In *Proceedings of the 14-th International Workshop on Spoken Dialogue Systems*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc.

Ze Yang, Wei Wu, Huang Hu, Can Xu, Wei Wang, and Zhoujun Li. 2021. Open domain dialogue generation with latent images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14239–14247.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152, Online. Association for Computational Linguistics.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv*.

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

A Prompts

The designed prompts for all descriptor-based approaches are shown as follows.

A.1 Desc - Summary

```
Please read the following dialogue context:
<dialogue_context>

Based on the dialogue context, please
summarize the information of speaker A.

Answers:
```

A.2 Desc - Guessing

```
Please read the following dialogue context:
<dialogue_context>

Based on the dialogue context, please describe
the photograph shared by speaker A.

Answers:
```

Method	VisDial				MMDialog			
	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg
FROMAGe (Koh et al., 2023)	20.8	44.9	56.0	40.57	-	-	-	-
Desc - Diag	11.00	24.03	32.56	22.53	22.10	34.30	41.00	32.47
Desc - Summary	19.43	39.05	48.84	35.77	14.60	20.00	22.50	19.03
Desc - Guessing	**29.46**	**53.73**	**64.68**	**49.29**	**31.30**	37.30	42.20	36.93
Desc - Queries	24.52	49.13	59.69	44.45	30.90	37.50	40.80	36.40
Desc - Queries (all)	24.3	47.34	59.74	43.80	31.10	**40.60**	**43.10**	**38.27**

Table 10: Retrieval performance for zero-shot on VisDial and MMDialog datasets (%). **Desc-Queries (all)** include the original five queries along with actions, atmosphere, or mood, and lighting queries We employ the LLaMA-2-13B-Chat with greedy decoding to ensure the correct format and reasoning capability.

A.3 Desc - Queries

```
Please read the following dialogue context:
<dialogue_context>

Based on the dialogue context, please describe
the photograph shared by speaker A.
List the answer in JSON format.
- main subject: {simply list the answer by ','}
- prominent objects in the foreground: {simply
list the answer by ','}
- background scene: {one background scene}
- events: {simply list the answer by ','}
- materials and attributes: {simply list the
answer by ','}

Answers:
```

B Detailed Results for Other Datasets

Table 10 presents detailed scores on the VisDial and MMDialog datasets using LLaMA-2-13B-Chat in zero-shot scenarios. It is worth noting that we sampled 1,000 instances from the MMDialog testing data for our experiments, limiting direct comparison with previous work due to the discrepancy of testing sets.