# Balancing Visual Context Understanding in Dialogue for Image Retrieval

**Zhaohui Wei[1], Lizi Liao[2], Xiaoyu Du[1]*, Xinguang Xiang[1]***
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology,
[2]Singapore Management University
zhwei@njust.edu.cn, lzliao@smu.edu.sg
duxy@njust.edu.cn, xgxiang@njust.edu.cn

## Abstract

In the realm of dialogue-to-image retrieval, the primary challenge is to fetch images from a pre-compiled database that accurately reflect the intent embedded within the dialogue history. Existing methods often overemphasize inter-modal alignment, neglecting the nuanced nature of conversational context. Dialogue histories are frequently cluttered with redundant information and often lack direct image descriptions, leading to a substantial disconnect between conversational content and visual representation. This study introduces VCU, a novel framework designed to enhance the comprehension of dialogue history and improve cross-modal matching for image retrieval. VCU leverages large language models (LLMs) to perform a two-step extraction process. It generates precise image-related descriptions from dialogues, while also enhancing visual representation by utilizing object-list texts associated with images. Additionally, auxiliary query collections are constructed to balance the matching process, thereby reducing bias in similarity computations. Experimental results demonstrate that VCU significantly outperforms baseline methods in dialogue-to-image retrieval tasks, highlighting its potential for practical application and effectiveness in bridging the gap between dialogue context and visual content.

## 1 Introduction

Recent advancements in communication technology have significantly enhanced online conversational systems (Jiang et al., 2019; Hosseini-Asl et al., 2020; Qin et al., 2023), which now play a crucial role in facilitating instant messaging and information sharing. However, purely textual exchanges often fail to fully convey the speaker's intentions and emotions, leading to the development of multimodal conversational systems that enrich dialogue with images and audio etc (Liao et al., 2018; Zhang et al., 2019; Tan and Bansal, 2019; Wang et al., 2020; Shuster et al., 2021; Yang et al., 2021; Ye et al., 2022). As shown in Figure 1, sharing photographs provides a vivid and intuitive means of communication, making image-sharing capabilities essential in dialogue systems.


Figure 1: An example of image-sharing in a multi-modal conversational system from the PhotoChat dataset.

Traditional text-to-image retrieval techniques (Faghri et al., 2018; Wang et al., 2019; Chun et al., 2021; Jia et al., 2021), which rely on direct image categories or detailed descriptions, fall short in dialogue contexts. The primary challenge is to select relevant images from a predefined repository based on the ongoing conversation, which involves understanding the dialogue content and accurately matching it with images. Previous studies (Zang et al., 2021; Yin et al., 2024) typically use dual-stream architectures for processing texts and images separately, followed by feature-based retrieval. Recent improvements in pre-trained visual-language models (Yin et al., 2024; Li et al., 2022, 2023a) have enhanced the accuracy of these systems by fine-tuning them for specific tasks.

However, current research (Zang et al., 2021;

*Corresponding Author.

Yin et al., 2024) often overemphasizes inter-modal alignment and overlooks the complexities of dialogue context. Dialogue context can be lengthy and cluttered with redundant information, such as greetings, small talk, and repeated phrases, thereby adding extra complexity to downstream retrieval tasks. Long texts make it more difficult for the model to understand, and may also be constrained by the encoder's limitation on the length of the input tokens. Consequently, it is crucial to better understand and extract image-related information from the dialogue context. In addition, dialogue context involves multi-turn exchanges, which may not directly describe the details of the images, leading to a significant disparity between the representations of dialogue contexts and images, directly affecting the matching results.

To address the aforementioned problems, we propose a systematic framework (**VCU**) for Balancing **V**isual **C**ontext **U**nderstanding in dialogue for image retrieval. Initially, to better understand the conversational content and mitigate interference from redundant information, we devise a two-step conversational content extraction process based on the Large Language Model (LLM). Specifically, in the first step, leveraging the comprehension and generative capabilities of LLM, we directly extract potential keywords related to images from the dialogue context. Then these keywords serve as hints for generating sentence-form visual description in the second step. Furthermore, to bridge the gap between the dialogue context and visual content, we enhance the visual representation using object-list texts associated with images. By computing similarity scores between token-level object-list text and patch-level image embeddings, we derive importance weights for each token, ultimately merging these weights with the original word embeddings and the global image embeddings to obtain enhanced image representation. Lastly, inspired by (Wang et al., 2024), we introduce a method to construct auxiliary query text collections to balance the matching, thereby reducing the bias during similarity calculation and improving retrieval accuracy.

To sum up, our contributions are threefold:

- We emphasize the importance of image descriptions driven by large language models and demonstrate their effectiveness in complex dialogue scenarios.
- We propose a framework that leverages LLMs for conversational context extraction, integrates object descriptions to enhance visual embeddings and constructs an auxiliary query text collection to balance matching.
- Comprehensive experiments show that our proposed VCU surpasses baselines, enabling precise image retrieval based on given conversational context.[1]

## 2 Related Work

### 2.1 Image-text Retrieval

The core of image-text retrieval tasks lies in effectively understanding and aligning the two distinct modalities of image and text. Early research, such as VSE++ (Faghri et al., 2018), proposes using the hardest negative triplet loss to learn superior joint visual-textual features. Subsequently, the focus of research has shifted to visual-semantic embedding (Wang et al., 2019; Chun et al., 2021; Jia et al., 2021). ALIGN (Jia et al., 2021) leverages a noisy dataset and employs contrastive learning to align visual and textual representations.

Moreover, several studies (Lee et al., 2018; Liu et al., 2019; Cui et al., 2021) have followed the development trend of attention mechanisms. SCAN (Lee et al., 2018) employs a cross-attention mechanism to establish finer-grained alignments between image regions and words, while BFAN (Liu et al., 2019) additionally considers the positional relationships among multiple image regions.

Recently, the emergence of large-scale pre-trained models has provided new perspectives for image-text retrieval tasks (Chen et al., 2020; Radford et al., 2021; Kim et al., 2021; Li et al., 2021; Bao et al., 2022; Li et al., 2022, 2023a). BLIP (Li et al., 2022) employs a multi-task learning strategy, thus excelling in various cross-modal tasks. In this study, our task can also be considered an image-text retrieval problem. However, unlike the traditional direct textual description of images, the textual component in our task comprises more complex dialogue histories.

### 2.2 Multimodal Dialogue

In recent years, research in multimodal dialogue has become a popular field, breaking through the limitations of traditional text-only interactions. Current research can be divided into two categories: the first one involves dialogue contexts that include

[1]The experimental codes are available at https://github.com/JupiterTop/VCU.

visual content, requiring models to better understand visual information or answer questions about the images (Tan and Bansal, 2019; Wang et al., 2020; Liao et al., 2021; Shuster et al., 2021; Park et al., 2021; Yang et al., 2021; Wu et al., 2022); the other one requires models to generate multimodal responses, which not only include texts but also involve retrieving or generating images (Agarwal et al., 2018; Sun et al., 2022, 2023; Yin et al., 2024).

PhotoChat (Zang et al., 2021) is the first open-domain multimodal dialogue dataset to facilitate the task of image-sharing. As the field evolves, an increasing number of studies (Shuster et al., 2020; Lee et al., 2021; Feng et al., 2023; Ahn et al., 2023; Lee et al., 2023) focus on this task, providing new available data. Most dialogue-to-image retrieval works use the raw dialogue history directly. DE* (Zang et al., 2021) uses a dual-encoder structure to encode dialogue history and visual content separately. PaCE (Li et al., 2023b) adopts a divide-and-conquer strategy to construct a pre-training framework suitable for different multimodal dialogue tasks. However, the dialogue context often contains redundant information that can complicate model understanding and may not directly describe the details of the image, adding complexity to downstream retrieval tasks. Therefore, our work focuses on better understanding and extracting image-relevant information from the dialogue context, thus bridging the gap between dialogue context and visual content.

## 3 Preliminary

**Task Formulation** For the task of dialogue-to-image retrieval, the model aims to select the image that best matches the current image-sharing intention from a pre-compiled database based on the dialogue history. We regard all the utterances preceding the image-sharing turn as dialogue history and denote it as $H = \{U_1, U_2, \cdots, U_{m-1}, U_m\}$, where $U_i$ represents the $i$-th turn of dialogue in the form of $\{$*User : Content*$\}$, and $m$ is the length of the dialogue history. At turn $m$+1, the model is required to retrieve an appropriate image from the pre-compiled image database $D = \{v_j, o_j\}_{j=1}^{N}$, where $v_j$ represents a candidate image and $o_j$ is the list of objects present in the corresponding image. Throughout the inference process, given the dialogue history $H$, the model will retrieve the image $v$ from image database $D$ for sharing.

**Pre-trained Encoders** CLIP (Radford et al., 2021) is a multi-modal model that efficiently processes textual and visual data. Pre-trained using a contrastive learning framework on a large dataset of text-image pairs, CLIP maps textual and visual embeddings into a common feature space, capturing semantic relationships between texts and images. In our research, we employ the text encoder $\psi_t$ and image encoder $\psi_v$ pre-trained by CLIP to extract embeddings of dialogue contexts and images. We then assess the alignment between these embeddings by calculating their cosine similarity. Our optimization objective is to enhance the model's performance in cross-modal retrieval task, specifically dialogue-to-image retrieval, by maximizing the similarity of correctly matched text-image pairs while minimizing the similarity of mismatched pairs.

## 4 Methodology

To address the challenges of conversational understanding and cross-modal matching in dialogue-to-image retrieval task, we propose a systematic framework VCU, as illustrated in Figure 2. The framework mainly consists of three main components: first, employing LLMs for conversational extraction in Section 4.1; second, using object-list texts to enhance visual embeddings in Section 4.2; and finally, balancing matching by building an auxiliary query text collection in Section 4.3. Furthermore, we detail our learning objectives in Section 4.4.

### 4.1 LLM-Driven Conversational Extraction

Dialogue history is typically lengthy and contains considerable redundant information, such as greetings, inquiries, etc. These redundant contents are often irrelevant to the image, so directly using the entire dialogue history as query text can easily lead to biases. Recent studies on large language models (LLMs), such as RAG (Lewis et al., 2020) and CoT (Wei et al., 2022), suggest that providing specific "hints" to input queries can guide LLMs to generate higher-quality content. Inspired by these findings, we propose a LLM-driven two-step conversational content extraction process to better understand the dialogue and refine the information relevant to retrieved image.

Specifically, given the dialogue history $H$, the first step involves using the LLM to directly select keywords $k$ related to the image to be shared, rather

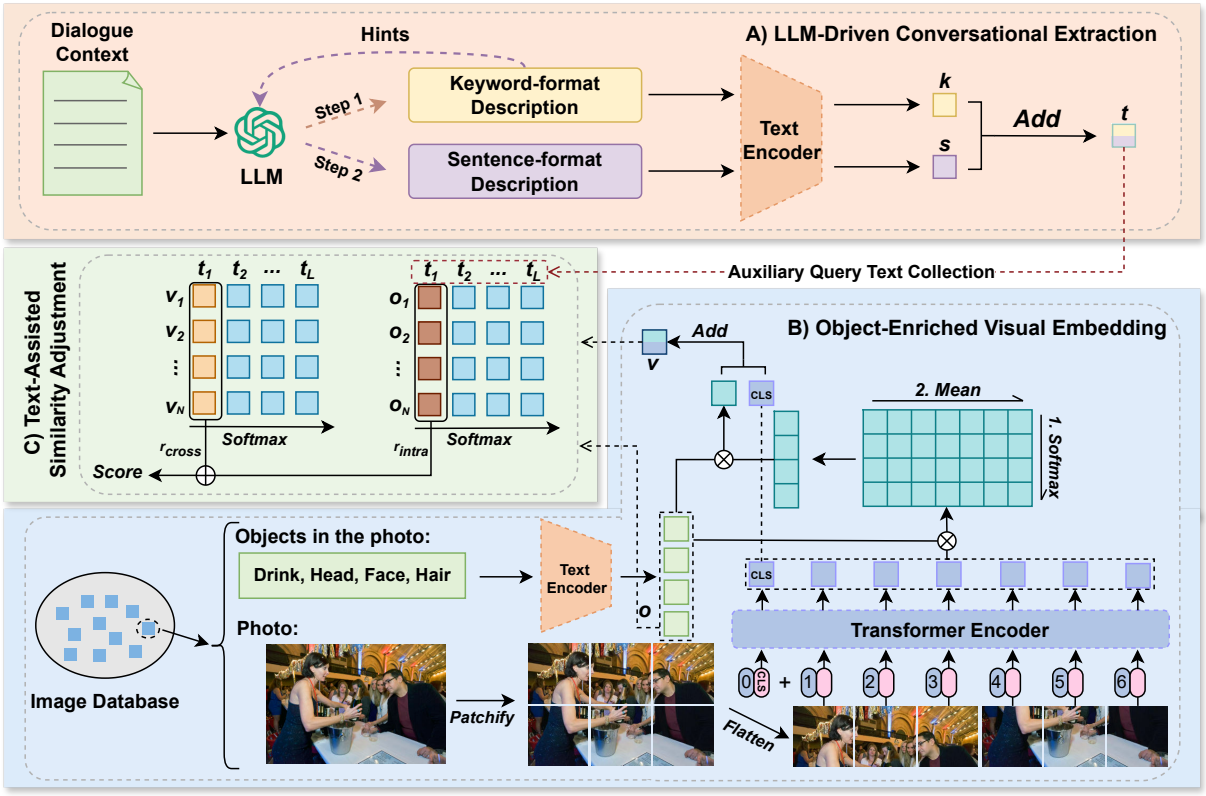


Figure 2: An overview of our **VCU** framework. **A)** Given the dialogue context, **VCU** employs the LLM to extract a keyword-format description of the image in step 1. In step 2, these keywords serve as input hints to generate a sentence-format visual description. Both descriptions are encoded to form dialogue context representation $\mathbf{t}$. **B)** We enrich the image representation with object-list texts. By computing the similarity scores between token-level object text and patch-level image embeddings, we derive the importance weights for each token. These weights are merged with the original word embeddings $\mathbf{o}$ and the global image embeddings to obtain enhanced image representation $\mathbf{v}$. **C) VCU** constructs an auxiliary query text collection from the dataset to balance the matching process. The final retrieval score is obtained by summing the cross-modal score $r_{cross}$ and the intra-modal score $r_{intra}$.

than generating new ones. In the second step, the keywords selected above serve as specific "hints" for input to generate descriptive text $s$ about the image in the form of sentence:

$$k = \text{LLM}([H]), \quad (1)$$

$$s = \text{LLM}([H, k]). \quad (2)$$

We believe that sentence-form texts contain more information than keywords alone, and generating sentences based on keywords can effectively guide the sentence-generation process. Ultimately, both forms of text are processed through the text encoder $\psi_t$ to obtain corresponding embedding representations. We add them as the final conversational representation for subsequent retrieval task:

$$\mathbf{k} = \psi_t(k), \mathbf{s} = \psi_t(s), \quad (3)$$

$$\mathbf{t} = \mathbf{k} + \mathbf{s}, \quad (4)$$

where $\mathbf{t}$ represents the representation of the dialogue history, which is obtained by adding the representations of keyword $\mathbf{k}$ and sentence $\mathbf{s}$. More details on LLM generation are provided in the Appendix A.1.

### 4.2 Object-Enriched Visual Embedding

During the description generation process, the shared images are invisible to the LLMs, so the quality of generated descriptions depends on dialogue history. In addition, dialogue history may not include statements that directly describe the details of images. These factors suggest that additional noises may be introduced during the generation process, resulting in a significant divergence between the dialogue context and the ground truth image, thereby affecting matching performance. To address this, we leverage resources on the image side, specifically utilizing object-list texts associated with images to enhance the visual representation. By emphasizing salient objects within the image and enhancing the image representation,

we aim to reduce the gap between the visual and textual representations.

To obtain more precise text-to-image saliency, we modify CLIP's basic encoders. For each pair of image and its corresponding object list text $\{v, o\}$ in the image database, we first using an adapted text encoder $\psi_t'$ to obtain token-level embeddings $\mathbf{O}^{M\times D} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_M}\}$ of the object list text $o$ as queries. For the image $v$, we perform patchify and flatten operations, add an additional classification token [CLS] and position embeddings, and then process them through the Transformer Encoder to obtain patch-level image embeddings $\mathbf{P}^{N\times D} = \{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_N}\}$, which serve as keys. By calculating the cosine similarity between the queries and keys, we have a similarity score matrix $\mathbf{W}^{M\times N}$, computed as

$$\mathbf{W} = \text{Softmax}(\mathbf{O}\mathbf{P}^T) = [s_{1,1}, \ldots, s_{M,N}], \quad (5)$$

$$s_{m,n} = \frac{\exp(\phi(\mathbf{w_m}, \mathbf{p_n}))}{\sum_{k=1}^{N} \exp(\phi(\mathbf{w_k}, \mathbf{p_n}))}, \quad (6)$$

where $m \in M, n \in N$ and each column of $\mathbf{W}^{M\times N}$ represents the normalized relevance scores between a particular patch and various tokens. $\phi(\cdot)$ denotes the cosine similarity function.

The saliency weight $\mathbf{w}_{token}$ of each token is calculated as the mean of its relevance scores with all patches. Finally, we compute the dot product between the obtained token-level weights and the original token embeddings $\mathbf{O}^{M\times D}$ and integrate the global image embeddings $\mathbf{p_{cls}}$ to derive the final enhanced visual representation $\mathbf{v}$:

$$\mathbf{v} = \mathbf{w}_{token} \cdot \mathbf{O} + \mathbf{p}_{cls}, \quad (7)$$

$$\mathbf{w}_{token} = [w_{token}(1), \ldots, w_{token}(M)], \quad (8)$$

$$w_{token}(i) = \frac{1}{N} \sum_{j=1}^{N} s_{i,j}. \quad (9)$$

### 4.3 Text-Assisted Similarity Adjustment

Although CLIP (Radford et al., 2021) is commonly used for cross-modal zero-shot learning, previous research (Wang et al., 2024) has found that the performance in text-to-image retrieval is affected by imbalances in similarity scores, leading to bias. Considering that there may be a significant gap between the predicted description texts and the actual image descriptions, we propose a text-assisted similarity adjustment method for dialogue-to-image retrieval to mitigate bias in dialogue understanding by balancing the matching process.

Specifically, when using conversational text $\mathbf{t}$ for image retrieval, we construct a set of text descriptions about the candidate images within the image database $D = \{v_i, o_i\}_{i=1}^{N}$, together with $\mathbf{t}$, forming an auxiliary query collection $\mathcal{S} = \{\mathbf{t_1}, \ldots, \mathbf{t_L}\} (L \leq N)$. Details of construction are provided in Section 5.4.2. We then calculate the similarity scores between each candidate image $\mathbf{v_i}$ and each query text $\mathbf{t_j}$ in the collection, normalizing these scores to balance the similarity score ranges between each image and different texts. Finally, we take the scores of candidate images and $\mathbf{t}$, which is adjusted by the query text set, as cross-modal scores $r_{cross}$:

$$r_{cross}(\mathbf{t}, \mathbf{v_i}) = \frac{\exp(\phi(\mathbf{t_j}, \mathbf{v_i}))}{\sum_{k=1}^{L} \exp(\phi(\mathbf{t_k}, \mathbf{v_i}))}. \quad (10)$$

During the matching process, these auxiliary texts offer diverse descriptions and perspectives, reducing the bias that may arise from a single text description (i.e., $\mathbf{t}$), thereby making the similarity scores more representative. Consequently, if a candidate image obtains relatively high similarity scores with $\mathbf{t}$ across multiple different descriptive texts, the match between the image and $\mathbf{t}$ is more probable to be accurate. In addition, we introduce intra-modal scores $r_{intra}$ between conversational text $\mathbf{t}$ and image object texts in the same way, and the sum of these two scores is used as the final retrieval score $\hat{r}$:

$$r_{intra}(\mathbf{t}, \mathbf{o_i}) = \frac{\exp(\phi(\mathbf{t_j}, \mathbf{o_i}))}{\sum_{k=1}^{L} \exp(\phi(\mathbf{t_k}, \mathbf{o_i}))}, \quad (11)$$

$$\hat{r} = r_{cross}(\mathbf{t}, \mathbf{v_i}) + r_{intra}(\mathbf{t}, \mathbf{o_i}). \quad (12)$$

### 4.4 Learning Objectives

During the training phase, we finetune CLIP using the train set to optimize the embedding space. To ensure stable model training and retain the pre-trained model's image features, we freeze the parameters of CLIP's image encoder $\psi_v$ and only adjust the parameters of the text encoder $\psi_t$. The text encoder is trained using a contrastive learning strategy and optimized with the widely-used symmetric cross-entropy loss function. Given a batch size of $B$, each training set triplet denotes $\{t, v, o\}$, where $t$, $v$, and $o$ respectively represent the dialogue context representation, the enhanced image representation, and the object-list text representation. The current triplet is treated as positive sample, while the remaining $B - 1$ triplets serve

as negative samples. Our optimization objective is to maximize the similarity between the conversational embeddings and both the image and object-list text embeddings within the positive samples, while minimizing the similarity for the negative samples. The overall loss function takes into account bidirectional optimization:

$$L_{t2v} = -\frac{1}{B}\sum_{k\in B}\log\frac{\exp(\phi(\mathbf{t_k},\mathbf{v_k})/\tau)}{\sum_{j\in B}\exp(\phi(\mathbf{t_k},\mathbf{v_j})/\tau)}, \quad (13)$$

$$L_{v2t} = -\frac{1}{B}\sum_{k\in B}\log\frac{\exp(\phi(\mathbf{t_k},\mathbf{v_k})/\tau)}{\sum_{j\in B}\exp(\phi(\mathbf{t_j},\mathbf{v_k})/\tau)}, \quad (14)$$

$$L_{total} = \frac{1}{4}(L_{t2v} + L_{t2v} + L_{t2o} + L_{o2t}), \quad (15)$$

where the loss functions of $L_{t2o}$ and $L_{o2t}$ are obtained by replacing image embedding $\mathbf{v}$ with object-list embedding $\mathbf{o}$ in Eq. 13 and Eq. 14.

## 5 Experiments

### 5.1 Experimental Setup

**Dataset.** We conduct experiments on two public datasets: **PhotoChat** (Zang et al., 2021) and **DialogCC** (Lee et al., 2023). **PhotoChat** is an open-domain multi-modal dialogue dataset manually constructed via a crowd-sourcing platform, and it is the first dataset to propose the image-sharing task. Each dialogue in PhotoChat includes one image shared, accompanied by a text describing the objects present in the image. **DialogCC**, on the other hand, is a high-quality multimodal dialogue dataset constructed through an automatic pipeline. Reflecting real-world scenarios and featuring diverse images, each dialogue in DialogCC comprises multiple image-sharing turns, with each turn involving several images. Each image is accompanied by a caption describing the objects. To maintain the complexity of dialogue history, we set all utterances preceding the final image-sharing turn as the dialogue history and designate the first image in the list as the ground truth. Detailed statistics are provided in Appendix A.2.

**Baselines.** We compare VCU with several robust baseline models, including: **VSE++** (Faghri et al., 2018), a cross-modal retrieval method that optimizes learning by leveraging hard negatives; **DE*** (Zang et al., 2021), a dual-encoder model that separately encodes textual and visual content; **PaCE** (Li et al., 2023b), a pre-training framework employing a divide-and-conquer strategy, suitable for various multimodal dialogue tasks; **CLIP** (Radford et al., 2021), a powerful pre-trained vision-language model that efficiently aligns text and image embeddings in the feature space and demonstrates exceptional zero-shot capability; and **DialCLIP** (Yin et al., 2024), a parameter-efficient prompt-tuning method designed specifically for multi-modal response retrieval tasks. In particular, we believe that conversational utterances closer to the image-sharing action are likely to be more relevant to the retrieved image. Consequently, when building the above **CLIP** baseline model, we opt to truncate the latter part of dialogue history exceeding the token limit of the text encoder, rather than the initial part. Similar to VCU, we consider the relationships between dialogue texts and images as well as between dialogue texts and object-list texts. We employ dot product to measure the similarity between embeddings.

**Implementation Details.** In our experiments, we primarily follow the dual-stream architecture based on CLIP, utilizing the pre-trained CLIP ViT-B/32. To ensure stability during model training, we freeze the parameters of the CLIP visual encoder and only finetune the text encoder. The model is trained on a single NVIDIA GeForce RTX 3090 GPU with the random seed fixed at 42 and a batch size of 56. We conduct one epoch of training on the train set and subsequently test on the test set. The Adam optimizer is employed with an initial learning rate of 1e-5. For the LLM in Section 1, we use Llama-3-8B-Instruct[2] and GPT-3.5-Turbo[3] for comparison. During the generation phase, we set the max-tokens to 80 while keeping other parameters at their default settings to ensure correct formatting and inference capabilities. Details of running time and memory consumption are provided in Appendix A.5.

**Evaluation Metrics.** We evaluate the models on the image retrieval aspect. Following previous works (Zang et al., 2021; Li et al., 2023b; Yin et al., 2024), we still employ **Recall@K** (**R@K**) as the evaluation metric to calculate the proportion of correctly retrieved images within the top K results. In particular, we select **R@1**, **R@5**, and **R@10** as metrics, and their cumulative sum, denoted as **R@Sum**, to provide a comprehensive assessment of the models' visual retrieval performance.

[2] https://github.com/meta-llama/llama3
[3] https://openai.com

| Methods | Training | PhotoChat | | | | DialogCC | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@Sum | R@1 | R@5 | R@10 | R@Sum |
| VSE++ | Required | 10.2 | 25.4 | 34.2 | 69.8 | - | - | - | - |
| DE* | Required | 9.0 | 26.4 | 35.7 | 71.1 | - | - | - | - |
| PaCE | Required | 15.2 | 36.7 | 49.6 | 101.5 | - | - | - | - |
| CLIP | Free | 20.2 | 37.3 | 45.7 | 103.2 | 4.1 | 12.3 | 18.4 | 34.8 |
| | Required | 27.0 | 50.0 | 60.2 | 137.2 | 12.8 | 34.2 | 47.2 | 94.2 |
| DialCLIP | Required | 19.5 | 44.0 | 55.8 | 119.3 | - | - | - | - |
| VCU(Llama) | Free | 35.6 | 57.6 | 65.9 | 159.1 | 12.6 | 32.2 | 43.2 | 88.0 |
| | Required | 40.2 | 62.9 | 71.1 | 174.2 | **14.3** | **36.6** | **49.1** | **100.0** |
| VCU(GPT) | Free | 37.6 | 58.8 | 67.1 | 163.5 | 11.0 | 29.2 | 39.7 | 79.9 |
| | Required | **42.8** | **64.0** | **73.4** | **180.2** | 13.5 | 35.0 | 46.1 | 94.6 |

Table 1: Main results of dialog-to-image retrieval on PhotoChat and DialogCC. Bold denotes the best result, and underline is the second-best result. "-": result is not available.

### 5.2 Main Results

In terms of dialog-to-image retrieval performance, we compare the proposed VCU with baseline methods on two datasets, with the main results summarized in Table 1. Overall, it is evident that our VCU significantly outperforms previous methods on all metrics. This demonstrates that our framework effectively captures critical visual information within complex dialogues, and improves cross-modal matching. Additionally, we observe that methods based on ViT image encoders (PaCE, CLIP, DialCLIP, and VCU), outperform those based on ResNet image encoders (VSE++ and DE*). This superiority can be attributed to ViT's advantages in modeling long-range dependencies, handling large-scale data, and extracting multi-scale features. Moreover, our VCU, similar to CLIP, supports zero-shot settings meaning it can achieve performance superior to baseline methods without extra training. The model's performance is also influenced by the language model (LLM) utilized. As shown in the last four rows of the table, in the PhotoChat dataset, the GPT-based VCU outperforms the Llama-based VCU, whereas in the DialogCC dataset, the reverse is true. This may be due to our setting about the DialogCC, which results in longer dialog contexts compared to PhotoChat. In such cases, GPT's robust generative capabilities might produce longer keyword descriptions than Llama, thereby introducing irrelevant information, generating inaccurate sentence-form descriptions, and affecting performance. It is worth noting that Llama-based generation requires additional manual data cleaning, as it often generates redundant content despite constraints in the prompt.

| Modules | | | PhotoChat | | |
|---|---|---|---|---|---|
| LCE | OVE | TSA | R@1 | R@5 | R@10 |
| × | × | × | 20.2 | 37.3 | 45.7 |
| ✓ | × | × | 31.4 | 51.9 | 60.1 |
| ✓ | ✓ | × | 31.5 | 53.2 | 61.4 |
| ✓ | × | ✓ | 37.2 | 57.5 | 66.7 |
| ✓ | ✓ | ✓ | **37.6** | **58.8** | **67.1** |

Table 2: Ablation study of three fundamental modules on PhotoChat under zero-shot GPT setting. Bold denotes the best result.

| Modules | | | DialogCC | | |
|---|---|---|---|---|---|
| LCE | OVE | TSA | R@1 | R@5 | R@10 |
| × | × | × | 4.1 | 12.3 | 18.4 |
| ✓ | × | × | 8.6 | 25.0 | 35.0 |
| ✓ | ✓ | × | 8.5 | 25.3 | 35.3 |
| ✓ | × | ✓ | 12.2 | 31.3 | 42.0 |
| ✓ | ✓ | ✓ | **12.6** | **32.2** | **43.2** |

Table 3: Ablation study of three fundamental modules on DialogCC under zero-shot Llama setting.

### 5.3 Ablation Study

In this section, we conduct the ablation study to evaluate the effectiveness of the VCU module under the zero-shot setting on the PhotoChat and DialogCC. The proposed VCU comprises three fundamental components: LLM-Driven Conversational Extraction (**LCE**), Object-Enriched Visual Embedding (**OVE**), and Text-Assisted Similarity Adjustment (**TSA**). As results shown in Table 2 and 3, both using raw CLIP (Row 1) as the baseline, we perform ablation experiments on these components. For the results on PhotoChat in Table 2, the introduction of LCE (Row 2) demonstrates

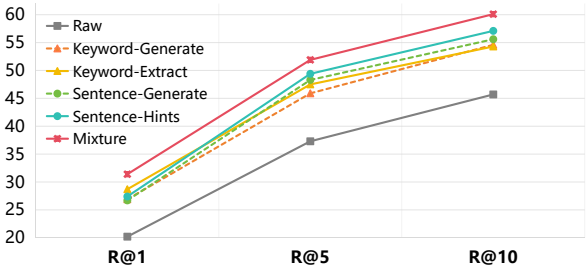


Figure 3: Effectiveness of LLM-Driven Extraction.

the capability to effectively extract key information from complex dialogues, showing improvements over the baseline by 11.2%, 14.6%, and 14.4% in R@1, R@5, and R@10 metrics, respectively. Subsequently, the individual introduction of OVE (Row 3) and TSA (Row 4) also resulted in performance enhancements to varying degrees, among which TSA notably improves the metric R@1 by 5.8%. Furthermore, introducing OVE and TSA together (Row 5) leads to further performance gains, indicating their synergy in reducing the modality gap and enhancing cross-modal matching, thereby significantly improving VCU's performance on PhotoChat. The ablation results on DialogCC have a similar trend.

### 5.4 Further Analysis

We conduct an in-depth analysis of VCU. The following experiments are performed on the test set of PhotoChat under zero-shot setting.

#### 5.4.1 Effectiveness of LLM-Driven Extraction

The dialogue history is often redundant and lacks specific image details, leading to modality gap in image retrieval. Therefore, it's crucial to effectively extract visual-related content from the dialogue for successful retrieval. We conduct extensive comparative experiments to evaluate the effectiveness of LLM-Driven Conversational Extraction. In these experiments, we ensured the use of the same LLM, GPT, and identical metric calculation method, differing only in generation methods. Used prompts are shown in Appendix A.1.

The pure CLIP (**Raw**) is employed as a baseline, utilizing the entire dialogue history as query text. The results shown in Figure 3 indicate that extracting keywords directly from the dialogue history (**Keyword-Extract**) yields better results than generating new keywords (**Keyword-Generate**). Furthermore, generating sentences based on hint keywords (**Sentence-Hints**) is more effective than generating sentences directly (**Sentence-Generate**), as the hint keywords effectively guide the direction of sentence generation. In addition, sentence-form descriptions outperform keyword-form descriptions, and using both forms together (**Mixture**) can achieve the best results. This aligns with the fact that sentence-form texts inherently contain more information than keyword-form texts.

| Strategies | Source | R@1 | R@5 | R@10 |
|---|---|---|---|---|
| Random | Train | 32.9 | 55.6 | 64.5 |
| Random | Test | 36.0 | 57.0 | 66.2 |
| Full | Test | **37.6** | **58.8** | **67.1** |

Table 4: Different strategies for selecting assistive texts on PhotoChat.

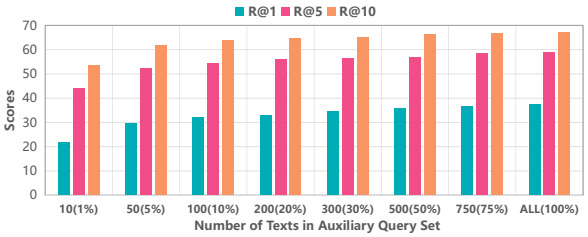


Figure 4: Comparison of results on PhotoChat under different settings about the number of texts within auxiliary query set.

#### 5.4.2 Strategies for Selecting Assistive Text

In the Text-Assisted Similarity Adjustment module, we construct an auxiliary query text collection $\mathcal{S}$. We further analyze how to select these texts. As shown in Table 4, we adopt two selection strategies: random and full. The random strategy involves randomly selecting a certain number of descriptive texts from the data source and ground truth text to form the set, while the full strategy uses all fixed number of descriptive texts within the test set. It is evident that regardless of whether the data source is from test set or train set, the full strategy significantly outperforms the random strategy, where the random quantity is half of number of images in the image database. Furthermore, we explore the impact of different numbers of query texts on retrieval performance under the random strategy in Figure 4. Specifically, as the number of texts in the auxiliary query collection increases, the balance matching effect improves. This is because more high-quality descriptions better differentiate the images. When candidate images match more suitable text in the auxiliary query collection than the current ground truth text, the influence on the target image naturally decreases, thereby improving the retrieval

performance of the ground truth text to the target image.

## 6 Conclusion

In this work, we proposed the VCU framework to enhance the understanding of conversational context and improve cross-modal matching in dialogue-to-image retrieval task. Based on the dialogue history, our approach leverages the LLMs for two-step extraction, enhances visual representation through object-list texts, and constructs auxiliary query collections to balance the matching. These effectively bridge the gap between dialogue context and visual content. Experimental results demonstrate that VCU outperforms baseline methods in image-sharing accuracy, highlighting the importance of considering the unique characteristics of dialogue. Future research could further explore better conversational extraction strategies and the more effective utilization of image-side information.

## Limitations

Our work is subject to the following limitations: (1) **Content generated by LLMs**. While our method effectively extracts conversational information, it does not perform any cleaning or filtering on the generated content, nor does it involve supervised training. This may lead to generated content being biased towards the dialogue side or mixed with wrong information, thereby weakening its relevance to image side. (2) **Dependence on data**. The proposed image enhancement method relies on datasets that include image captions or object-list descriptions. This imposes certain pre-processing requirements on the datasets, involving the use of techniques such as object detection or caption generation. (3) **Analysis of dialogue history**. We plan to further dissect the dialogue history content and analyze sentence structure. By employing more fine-grained methods to eliminate redundant information within the dialogue, we aim to enhance the accuracy of predicting visual content.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62172226, Grant 62272230, and the 2021 Jiangsu Shuangchuang (Mass Innovation and Entrepreneurship) Talent Program (JSSCBS20210200).

## References

Shubham Agarwal, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Improving context modelling in multimodal dialogue generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 129–134.

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. Mpchat: Towards multimodal persona-grounded conversation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424.

Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. 2021. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 797–806.

Fartash Faghri, DavidJ. Fleet, Jamie Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. *British Machine Vision Conference*.

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.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

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.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. In *The world wide web conference*, pages 2879–2885.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

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.

Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyon Myaeng. 2021. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 897–906, Online. Association for Computational Linguistics.

Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, Jonghwan Hyeon, and Ho-Jin Choi. 2023. Dialogcc: An automated pipeline for creating high-quality multimodal dialogue datasets. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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. In *International conference on machine learning*, pages 19730–19742. PMLR.

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.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

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.

Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. Mmconv: an environment for multimodal conversational search across multiple domains. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 675–684.

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.

Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM international conference on multimedia*, pages 3–11.

Sungjin Park, Taesun Whang, Yeochan Yoon, and Heuiseok Lim. 2021. Multi-view attention network for visual dialog. *Applied Sciences*, 11(7):3009.

Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5925–5941.

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, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429.

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.

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.

Xiaowen Sun, Jiazhan Feng, Yuxuan Wang, Yuxuan Lai, Xingyu Shen, and Dongyan Zhao. 2023. Teaching text-to-image models to communicate. *arXiv preprint arXiv:2309.15516*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on*

*Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Hanyao Wang, Yibing Zhan, Liu Liu, Liang Ding, and Jun Yu. 2024. Balanced similarity with auxiliary prompts: Towards alleviating text-to-image retrieval bias for clip in zero-shot learning. *arXiv preprint arXiv:2402.18400*.

Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position focused attention network for image-text matching. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3792–3798.

Yue Wang, Shafiq Joty, Michael Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. 2020. Vd-bert: A unified vision and dialog transformer with bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3325–3338.

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

Yuxia Wu, Lizi Liao, Gangyi Zhang, Wenqiang Lei, Guoshuai Zhao, Xueming Qian, and Tat-Seng Chua. 2022. State graph reasoning for multimodal conversational recommendation. *IEEE Transactions on Multimedia*, 25:3113–3124.

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.

Chenchen Ye, Lizi Liao, Suyu Liu, and Tat-Seng Chua. 2022. Reflecting on experiences for response generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5265–5273.

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*.

Zheng Zhang, Lizi Liao, Minlie Huang, Xiaoyan Zhu, and Tat-Seng Chua. 2019. Neural multimodal belief tracker with adaptive attention for dialogue systems. In *The world wide web conference*, pages 2401–2412.

# A Appendix

## A.1 Prompts for LLMs

Here, we show the prompts used in the module of LLM-Driven Conversational Extraction, as shown in Figure 5. For the two different LLMs, GPT-3.5 and Llama-3, due to their different generation capabilities, we have made minor adjustment in the design of prompts to ensure that the generated content can be presented in a correct and uniform form. Specifically, when using Llama-3 to generate keywords based on the dialogue context, an additional restriction of "Output only keywords, separated by ," is added.

The prompts utilized in the comparative experiment are presented in Table 5.

## A.2 Statistics of Datasets

We conduct a comprehensive analysis of the datasets used in our experiments. PhotoChat (Zang et al., 2021) is a public multimodal dialogue dataset, divided into training, validation, and test sets in a ratio of 10:1:1, containing a total of 10,917 different images. DialogCC (Lee et al., 2023) aims to exploit the capabilities of the Large Language Model (LLM) and Vision Language Model (VLM) to automatically construct a high-quality multimodal dialogue dataset, encompassing 83,370 dialogues. Unlike PhotoChat, where only one image is shared per dialogue, DialogCC dialogues contain multiple image-sharing turns, with each turn containing multiple images. We standardize the DialogCC setting to make it suitable for dialogue-to-image retrieval tasks by setting all utterances before the last image-sharing turn as dialogue history and the first image in the sharing list as ground truth. In addition, we filter out dialogues containing images that cannot be accessed and downloaded via Python scripts based on their URLs. Detailed statistics of the filtered data are presented in Table 6.

## A.3 Additional analysis

### A.3.1 Differences Between LLMs for Extraction

To further investigate the specific impact of different LLMs on the LLM-Driven Conversational Extraction module in Section 4.1, we conduct additional experimental explorations on this module independently.

As illustrated in Figure 6 and Figure 7, the quality of the descriptions they generate also varies. In particular, GPT-3.5 always shows an advantage

| Category | Description | Prompt |
|---|---|---|
| Keyword-Generate | Directly predict and generate keywords related to the image from the dialogue history | Please read the following dialogue context: [context]. Based on the dialogue context, please use some short keywords to describe the objects or events that may appear in the photograph shared by speaker A. Answers: |
| Sentence-Generate | Generate a single descriptive sentence about the image using the LLM | Please read the following dialogue context: [context]. Based on the dialogue context, please use one sentence to describe the objects or events that may appear in the photograph shared by speaker A. Answers: |

Table 5: Prompts utilized in the comparative experiment for Section 4.1.

| Dataset | Split | Raw Images | Filtered Images | Raw Dials | Filtered Dials |
|---|---|---|---|---|---|
| PhotoChat | train | 8,917 | 8,878 | 10,286 | 10,275 |
| | valid | 1,000 | 998 | 1,000 | 998 |
| | test | 1,000 | 1,000 | 1,000 | 1,000 |
| DialogCC | train | 30,253 | 26,636 | 68,402 | 60,480 |
| | valid | 4,875 | 4,508 | 7,644 | 6,869 |
| | test | 4,809 | 4,571 | 7,324 | 6,993 |

Table 6: Detailed statistics of PhotoChat and DialogCC after filtering unavailable images and corresponding dialogues.

over Llama-3 in terms of description generation on PhotoChat, regardless of the generation method used. On the DialogCC dataset, however, the situation is exactly the opposite, with Llama-3 performing better. This difference may be due to the complexity and length of the dialogue history in DialogCC. Since GPT-3.5 has better generation capabilities, it may be able to extract more keywords, but it may also introduce more wrong information, which in turn affects its performance on this dataset.

#### A.3.2 Impact of Intra-Modal Score

In the process of calculating retrieval scores, in addition to utilizing the basic cross-modal scores $r_{cross}$ between dialogue text and candidate images, we also incorporate intra-modal scores $r_{intra}$ between dialogue text and object-list texts. We assess the impact of the intra-modal scores through ablation studies on PhotoChat and DialogCC datasets. The results presented in Table 7 demonstrate that the inclusion of intra-modal scores significantly enhances recall performance, leading to substantial improvements.

### A.4 Case Study

In Figures 8 and 9, we conduct a case study to compare the image retrieval performance of our proposed VCU with the baseline CLIP model.

Specifically, the first example in Figure 8 focuses on the zero-shot experimental setting on the PhotoChat test set, where VCU uses GPT-3.5 as the LLM. The experimental results show that our model effectively exploits the robustness of the LLM to accurately extract keywords related to the image content from the conversational context and generate accurate image sentence descriptions based on these keywords. Finally, the ground truth image is successfully retrieved using the text description as a query, while the CLIP model fails to achieve successful retrieval. Further in-depth analysis shows that our model not only captures the keyword "*cake*", but also captures the keyword "*servant*" very well, so that the retrieved images are more related to the person. In contrast, the images retrieved by the CLIP model are only related to

**Step One:**

**Prompt**

(GPT-3.5) Please read the following dialogue context: `[Context]`.
Based on the dialogue context, please use some short keywords to describe the objects or events that may appear in the photograph shared by speaker A. You MUST select the keywords in the given dialogue context, NOT generate a new keyword. Answers:

(Llama-3) Please read the following dialogue context: `[Context]`.
Based on the dialogue context, please use some short keywords to describe the objects or events that may appear in the photograph shared by speaker A. You MUST select the keywords in the given dialogue context, NOT generate a new keyword.
Output only keywords, separated by ','. Answers:

**Answer**

`[Keywords]`

---

**Step Two:**

**Prompt**

(GPT-3.5 & Llama-3) Please read the following dialogue context: `[Context]`.
Based on the dialogue context, please use one sentence to describe the objects or events that may appear in the photograph shared by speaker A.
Here are some keywords as hints: `[Keywords]`. Answers:

**Answer**

`[Sentence]`

Figure 5: Used prompts for GPT-3.5 and Llama-3.

| Setting | PhotoChat | | | | DialogCC | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@Sum | R@1 | R@5 | R@10 | R@Sum |
| w/o rintra | 36.9 | 57.9 | **67.7** | 162.5 | 11.9 | 30.8 | 41.5 | 84.2 |
| w/ rintra | **37.6** | **58.8** | 67.1 | **163.5** | **12.6** | **32.2** | **43.2** | **88.0** |

Table 7: Comparison of zero-shot results with/without intra-modal score $r_{intra}$. Results on PhotoChat use GPT-3.5 as LLM, while those on DialogCC employ Llama-3.

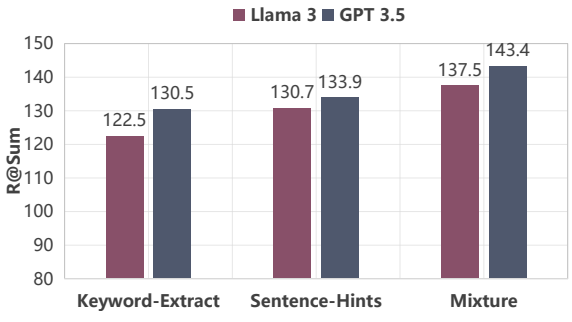


Figure 6: Differences in extraction between Llama-3 and GPT-3.5 on PhotoChat.

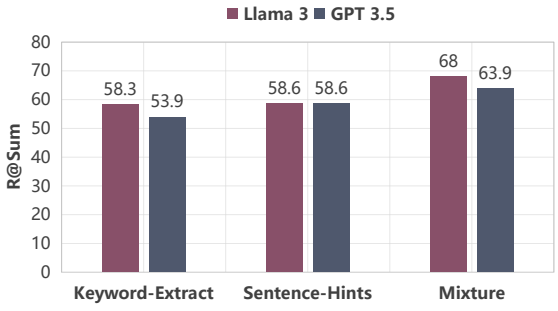


Figure 7: Differences in extraction between Llama-3 and GPT-3.5 on DialogCC.

"*cake*".

The second case in Figure 9 examines the performance of the models on the DialogCC test set, also under the zero-shot setting, but VCU uses Llama-3 as the LLM. Similarly, the experimental results show that VCU shows a higher concentration in retrieving image content, mainly focusing on house images related to "*balcony*". The image content

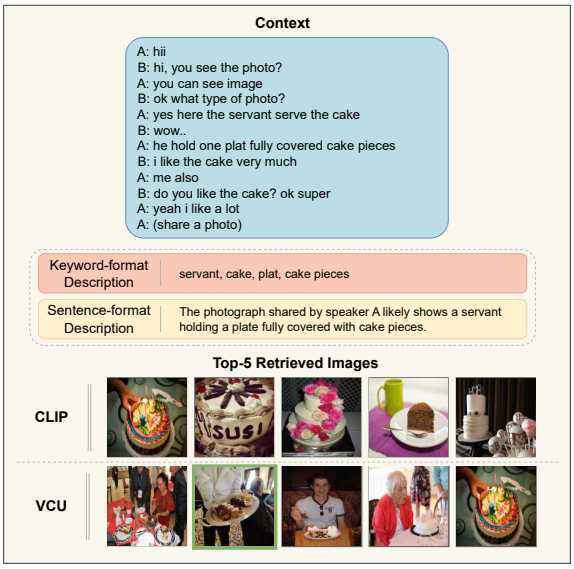


Figure 8: An example of results on PhotoChat dataset about image retrieval performance between VCU (used GPT-3.5) and CLIP.


Figure 9: An example of results on DialogCC dataset about image retrieval performance between VCU (used Llama-3) and CLIP.

retrieved by the CLIP model is quite different, including images related to "*beach*" and "*family*". Although both models successfully retrieved the ground truth image within the top 5 images, VCU retrieved a higher ranking, which will have a positive impact on the improvement of other indicators such as "Recall@1".

In summary, through these two case studies, we can clearly see that our proposed VCU shows a significant advantage in image retrieval performance compared to the baseline CLIP model.

| Dataset | GPT-3.5 | | Llama-3 | |
|---|---|---|---|---|
| | KEY | SEN | KEY | SEN |
| PhotoChat | 2.22 | 2.46 | 0.66 | 0.90 |
| DialogCC | 3.34 | 4.45 | 0.66 | 1.00 |

Table 8: Comparison of the time required to generate examples in PhotoChat and DialogCC using GPT-3.5 and Llama-3, measured in seconds. **KEY** and **SEN** represent the Keyword-format and Sentence-format description generation stages respectively.

### A.5 Running Time and Memory Consumption

In this study, we sequentially executed the generation of Llama-3-8B-Instruct and the CLIP training phase on a single NVIDIA RTX 3090, while we made API calls to utilize GPT-3.5 for generation. We provide detailed generation time in Table 8. On the PhotoChat standard test set (1,000 samples), the average generation time per sample for GPT-3.5 is 4.68 seconds, and for Llama-3, it is 1.56 seconds. The average time taken over whole test set for balanced similarity computation is 49 seconds, consuming 2GB of memory. For the DialogCC standard test set (6,993 samples), the average generation time per sample for GPT-3.5 is 7.79 seconds (the difference from the time reported in PhotoChat mainly because of the different time period for api calls and the prompt length), and for Llama-3, it is 1.66 seconds. The balanced similarity computation costs 4 minutes 14 seconds and 2GB of memory. Moreover, the average memory consumption for Llama-3-8B-Instruct during running is 16GB.