MsRAG: Knowledge Augumented Image Captioning with Object-level Multi-source RAG

Yuming Qiao[1] , **Yuechen Wang**[1] , **Dan Meng**[1,*] , **Haonan Lu**[2] , **Zhenyu Yang**[2] , **Xudong Zhang**[1]

[1]OPPO Research Institute
[2]OPPO AI Center

qym.2021@tsinghua.org.cn, wyc9725@mail.ustc.edu.cn, mengdan90@163.com, {luhaonan, yangzhenyu, zhangxudong}@oppo.com

Abstract

Language-Visual Large Models (LVLMs) have made significant strides in enhancing visual understanding capabilities. However, these models often struggle with knowledge-based visual tasks due to constrains in their pre-training data scope and timeliness. Existing Retrieval-Augmented Generation (RAG) methods can effectively solve the problem but primarily rely on user queries, limiting their applicability in scenarios without explicit language input. To overcome these challenges, we introduce **MsRAG**, a knowledge-augmented captioning framework designed to effectively retrieve and utilize external ***real-world knowledge***, particularly in the absence of user queries, and perform dense captioning for subjects. **MsRAG** comprises three key components: (1) **Parallel Visual Search Module**. It retrieves fine-grained object-level knowledge using both online visual search engines and offline domain-knowledge databases, enhancing the robustness and richness of retrieved information. (2) **Prompt Templates Pool**. The prompt pool dynamically assigns appropriate prompts based on retrieved information, optimizing LVLMs' ability to leverage relevant knowledge under complex RAG conditions. (3) **Visual-RAG Alignment Module**, which employs a novel visual prompting method to bridge the modality gap between textual RAG content and corresponding visual objects, enabling precise alignment of visual elements with their text-format RAG content. To validate the effectiveness of **MsRAG**, we conducted a series of qualitative and quantitative experiments. The evaluation results demonstrate the superiority of **MsRAG** over other methods.

1 Introduction

The rise of Large Language Models (LLMs) has significantly enhanced the utility of Artificial Intelligence Generated Content (AIGC) [Touvron *et al.*, 2023; Bai *et al.*, 2023a], drawing extensive attention and research interest from both academia and industry. Leveraging the robust natural language understanding capabilities of foundation LLMs, the emergence of Language-Visual Large Models (LVLMs) has notably bridged the gap between textual and visual modalities, greatly improving models' vision comprehension abilities [Liu *et al.*, 2023; Bai *et al.*, 2023b]. Traditional benchmarks for text-image understanding, such as NLVR2 Test [Suhr and Artzi, 2019], CommercialAdsDataset[Zhu *et al.*, 2022], nocaps [Agrawal *et al.*, 2019], and VQA [Goyal *et al.*, 2017], which focus on visual reasoning, image captioning, and simple visual question answering (VQA) tasks, are no longer sufficient to comprehensively assess LVLMs visual understanding capabilities. Consequently, more complex benchmarks like knowledge-based VQA tasks [Marino *et al.*, 2019; XENOS *et al.*, 2023] have been proposed to test the multimodal understanding capabilities of LVLMs. However, constrained by the timeliness and knowledge breadth of pretraining data, these models often exhibit hallucination issues in ***real-world knowledge*** based VQA tasks [Bai *et al.*, 2024].

DPO [Rafailov *et al.*, 2024], RAG [Lewis *et al.*, 2020], etc. are currently effective solutions for addressing model hallucinations. Among them, RAG usually provides the model with information related to user queries or images through incontext learning, namely incorporating real-world knowledge as prior information input to the model. This approach can effectively reduce model hallucinations, improving response quality of visual tasks (such as visual understanding, visual reasoning, and captioning) without post-training. Consequently, the RAG methods are widely used in LLMs/LVLMs.

Similar to visual reasoning and VQA tasks, image captioning plays an important role in LVLMs based visual tasks, and high quality captioning data is critical for enhancing the performance of visual tasks. The pre-training phase of current LVLMs heavily relies on image captioning tasks to achieve effective visual and text modality alignment [McKinzie *et al.*, 2025]. The quality of the captioning data also greatly determines the knowledge boundary of the foundation model, impacting overall performance [Yin *et al.*, 2023]. Knowledge-augmented image captioning [Yu *et al.*, 2024; Jiang *et al.*, 2024] can effectively combines image descriptive language and real-world knowledge, which is crucial for practical applications or synthesizing high-quality data for

---

** Corresponding author.

In this paper, we formally define real-world knowledge to include time-sensitive knowledge and domain-specific knowledge.

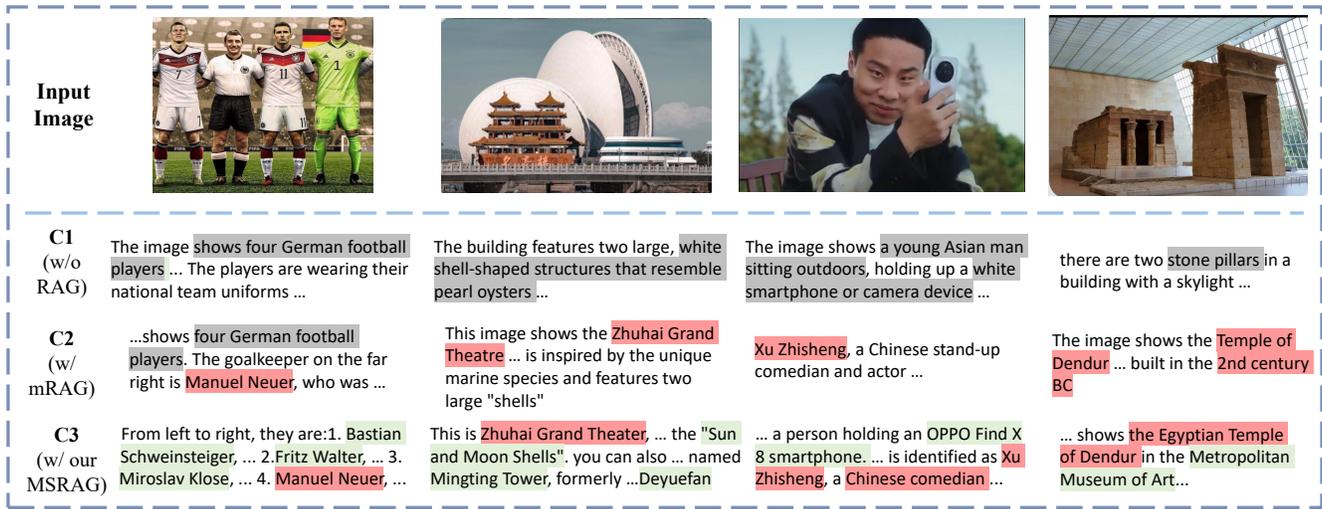|				
Input Image				
C1 (w/o RAG)	The image shows four German football players … The players are wearing their national team uniforms …	The building features two large, white shell-shaped structures that resemble pearl oysters …	The image shows a young Asian man sitting outdoors, holding up a white smartphone or camera device …	there are two stone pillars in a building with a skylight …
C2 (w/ mRAG)	…shows four German football players. The goalkeeper on the far right is Manuel Neuer, who was …	This image shows the Zhuhai Grand Theatre … is inspired by the unique marine species and features two large "shells"	Xu Zhisheng, a Chinese stand-up comedian and actor …	The image shows the Temple of Dendur … built in the 2nd century BC
C3 (w/ our MSRAG)	From left to right, they are:1. Bastian Schweinsteiger, … 2.Fritz Walter, … 3. Miroslav Klose, … 4. Manuel Neuer, …	This is Zhuhai Grand Theater, … the "Sun and Moon Shells". you can also … named Mingting Tower, formerly …Deyuefan	… a person holding an OPPO Find X 8 smartphone. … is identified as Xu Zhisheng, a Chinese comedian …	… shows the Egyptian Temple of Dendur in the Metropolitan Museum of Art…

Figure 1: Some captioning results of our MsRAG compared to other methods. The proposed MsRAG enables LVLMs achieving high-quality knowledge augmented captioning in various domains. Text in green represents knowledge gain of MsRAG. Text in red is retrieved contents of mRAG. Text in grey is common descriptive language of the image.

improving visual comprehension capabilities.

In short, we take advantages of both RAG and knowledge-augmented image captioning technologies to equip LVLMs with up-to-date and domain-specific knowledge, thereby enhancing the performance of image captioning tasks. Unfortunately, current mainstream RAG solutions are tailored for scenarios with user queries [Gao *et al.*, 2023; Zhao *et al.*, 2023], making them unsuitable for situations without such queries. Moreover, interpreting user intent and effectively utilizing retrieved content become challenging in the absence of user queries. To address this issue, we propose the MsRAG framework to guide LVLMs in effectively retrieving and utilizing real-world knowledge without user queries (see Fig. 1). By doing so, it makes AIGC technology more accessible to everyday users and unlock its full potential.

The MsRAG framework aims to solve two key research challenges. (1) **Multi-source RAG information acquisition**. How can abundant and related RAG content be obtained without user queries? (2) **Complex RAG content utilization**. How can diverse RAG information be effectively used, while noisy information can be automatically ignored and LVLMs' self-correction capabilities can be improved? To evaluate the effectiveness of MsRAG, we integrate several mainstream LVLMs, including GPT4o, Claude-sonnet-3.5, Qwen2VL [Bai *et al.*, 2023b], and InternVL2 [Chen *et al.*, 2024], into the framework. We also introduce a **k**nowledge-**a**ugmented image **c**aptioning benchmark KAC-dataset that covers multiple domains. By assessing the captioning results, we tested the MsRAG's ability to retrieve and utilize RAG information under non-query conditions. We further evaluated our MsRAG on existing benchmarks like CapFusion [Yu *et al.*, 2024] and Kale [Awadalla *et al.*, 2024]. Results show that MsRAG significantly improves the knowledge-based captioning capabilities of LVLMs, demonstrating its effectiveness in enhancing real-world and domain-specific knowledge integration and utilization. Additionally, MsRAG demon-

strates robustness by filtering out irrelevant or noisy RAG information and correcting inaccuracies, thereby improving the overall quality of generated captions. This proves the successful fusion of RAG and knowledge-augmented image captioning techniques, leading to notable improvements in object-level knowledge based image captioning tasks.

Our motivation is inspired by the reliance of current QA systems on user language queries. With the willing of effectively applying RAG system to image captioning task, we propose MsRAG, an effective RAG framework that can address the knowledge gap of existing RAG systems in knowledge-augmented image captioning tasks. Our contributions can be summarized as follows.

- We propose a general tuning-free RAG framework to boost LVLMs' knowledge-based image captioning capability. MsRAG is plug-and-play that can be easily, flexibly, and cost-effectively integrated with any open-source or close-source LVLMs.

- We develop an object-level Parallel Visual Search module to enrich retrieved information from both online visual search engines and offline domain-specific knowledge databases.

- We propose a Visual-RAG alignment module to bridge the gap between object-level image semantics and their text-format RAG contents.

2 Related Work

2.1 Large Vision-Language Models (LVLMs)

The significant improvement in natural language understanding by LLMs has established a strong foundation for advanced multi-modal comprehension. Prominent mainstream LVLMs, such as Qwen2VL [Bai *et al.*, 2023b], InternVL [Chen *et al.*, 2024], and LLaVA [Liu *et al.*, 2023], are typically constructed using pre-trained LLMs and visual encoders. These LVLMs demonstrate excellent performance

in comprehending both user instructions and image content through visual-text modality alignment training. They deliver impressive performance in diverse visual understanding tasks. Nevertheless, due to constraints in the timeliness and breadth of pre-training corpus, LVLMs frequently exhibit inaccuracies when confronted with scenarios requiring knowledge beyond their initial training scope.

2.2 Retrieval-Augmented Generation (RAG)

RAG is widely used in LLMs to mitigate model hallucination. Typically, it takes a user query as input, retrieves relevant external information, and enhances the output with the retrieved content [Gao *et al.*, 2023]. This technique has also been applied to LVLMs for knowledge-based VQA tasks. For example, Wiki-LLaVA [Caffagni *et al.*, 2024] employs a two-stage retrieval process, first using the input image and then refining the content with the user query. EchoSight [Yan and Xie, 2024] introduces a Q-Former based reranker for retrieval using both query and image inputs, while SearchLVLMs [Li *et al.*, 2024] enhance queries before retrieving and filtering web content. Entity-centric methods such as SnapNTell [Qiu *et al.*, 2024], MAR [Zhang *et al.*, 2024b], and MuKA [Deng *et al.*, 2025] leverage visual grounding to retrieve knowledge about specific objects in images.

All approaches mentioned above rely on extra expert models to handle the retrieved content and ignore the inherent capacity of the LVLMs. Self-RAG [Asai *et al.*,] strengthens the model's self-evaluation capability by setting special reflection tokens during training. Inspired by self-RAG, mR^{2}AG [Zhang *et al.*, 2024a] introduces a retrieval-reflection mechanism guiding the model to step-by-step judge the relevance of the retrieved content with input image and user query, reranking the retrieved content based on its own capacity. However, these RAG-based methods mainly focus on VQA scenarios that rely on user queries to retrieve domain-specific and time-sensitive knowledge. Few research addresses the retrieval and utilization of real-world knowledge under non-query conditions. This gap highlights the need for new approaches that can effectively generate image captioning without explicit user queries, expanding the applicability of such systems to a broader range of scenarios.

2.3 Knowledge Augmented Image Captioning

Yu [Yu *et al.*, 2024] analyzed the impact of captioning data in the pre-training phase on the model's knowledge boundary, finding that LVLMs trained based on existing synthetic captioning data such as LAION-COCO [Schuhmann *et al.*, 2022] would experience a significant degree of knowledge loss. This is attributed to the fact that the captions output by the captioning model used by LAION-COCO, such as BLIP [Li *et al.*, 2022], often replace some specific content with more common placeholders, resulting in the final trained LLMs tends to caption image with simple language structures containing only basic semantic information rather than detailed object-level real-world knowledge.

To address this issue, CapFusion [Yu *et al.*, 2024] generates knowledge-augmented image captioning data by prompting GPT with raw captions from Laion-2B and descriptive captions from LAION-COCO, and the experimental re-

sults prove that such captioning data can effectively broaden the model's knowledge boundary. Inspired by CapFusion, BLIP3-KALE [Jiang *et al.*, 2024] further improves the granularity and efficiency of data annotation with a two-stage process. First, LVLM was adopted to create an initial pool of knowledge augmented captioning data. Then, captions from stage 1 were used to distill a light-weight VLM to accelerate the captioning process. Compared with previous captioning tasks, knowledge-augmented image captioning is more valuable in both real-world image caption applications and high quality captioning data generator for training stronger LVLMs. However, current knowledge-augmented image captioning task primarily focuses on extending existing datasets, and does not involve open-domain images and the application of RAG systems to achieve general knowledge-augmented image captioning. In this paper, We propose **MsRAG**, focusing on leveraging LVLMs to effectively retrieve and utilize fine-grained object-level multi-source RAG information under non-query conditions, so as to achieve high-quality, generalized knowledge-augmented image captioning.

3 MsRAG Framework

In this section, we introduce our MsRAG, a tuning-free framework enabling LVLMs to achieve knowledge augmented image captioning with object-level multi-source RAG contents. MsRAG framework is shown in Fig. 2, given an input image I without language query, we focus on investigating how to retrieve and utilize real-world knowledge (e.g., domain-specific knowledge) without user query, and incorporate retrieved knowledge into captioning results.

For the input image I, we propose a parallel visual search module containing a pre-process module and an RAG content summary module. In the pre-process phase, an object detection model is applied to identify objects in the image and crop object-aware image I_o. Subsequently, I_o is fed into the RAG content summary module to further extract and analyze detailed object-level knowledge from both online visual search engines and offline domain-specific knowledge databases. Based on the parsed information output from the RAG content summary module, we designed a set of prompt templates. This pool dynamically selects the appropriate prompt according to the object detection results and RAG contents, enhancing the LVLMs' ability to utilize complex RAG information effectively. We also propose a Visual-RAG alignment module (VRAM) to align the object-aware image I_o with its corresponding text-format RAG content $\{R_{ol}^i\}_{i=1}^n$. This process generates a marked image I_o^m and an aligned prompt P_A, allowing LVLMs to better understand the relationship between visual objects and their associated object-level knowledge. Finally, the marked image I_o^m and the Visual RAG aligned prompt P_A are sent to the LVLMs to generate knowledge-augmented image captions.

3.1 Parallel Visual Search Module

Given an input image I, undergoing preprocessing module, an object-aware image I_o is obtained. We crop different objects according to their bounding boxes and then merge & filter them based on their entity attributes to obtain sub-images
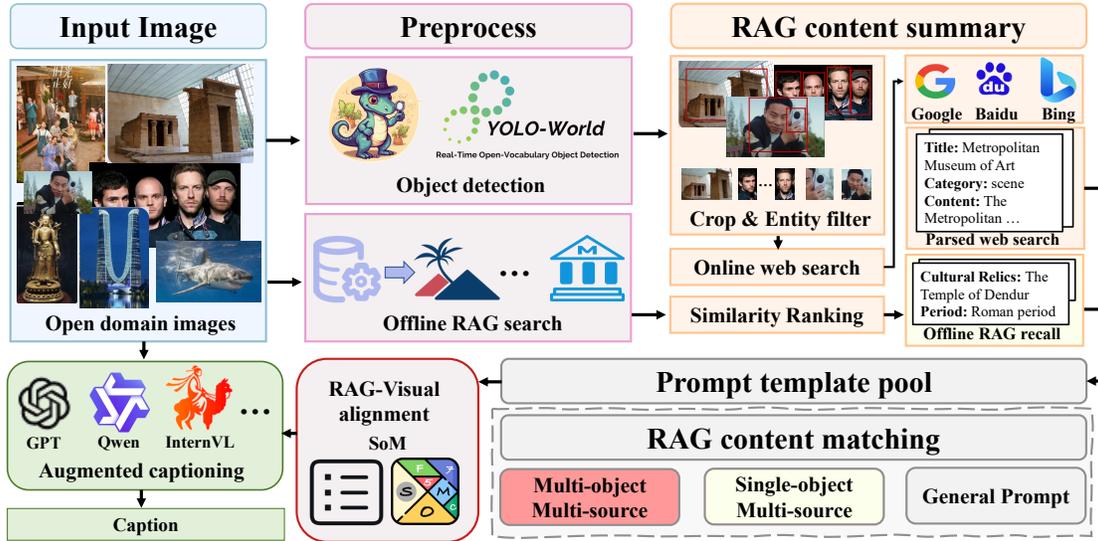
Figure 2: Overview of our propsed MsRAG, a training-free framework allows LVLMs to caption image with real-word knowledge.

I_o^i, i means the i-th object in the image. These sub-images are then inputted into online visual search engines to obtain object-level knowledge $\{R_{ol}^i\}_{i=1}^n$ from the websites. However, online visual search engines perform well in retrieving time-sensitive knowledge but fail to provide accurate information in specific domains like culture relics, products, and scenes. We organized an offline domain-knowledge database to compensate for the domain-specific knowledge gap in web search. Similarly, we can also retrieve offline object-level knowledge $\{R_{of}^i\}_{i=1}^n$ from offline database. For the retrieval information of each object from both online R_{ol}^i and offline R_{of}^i, we select the top three search results $[R_{ol}^{i,1:3}, R_{of}^{i,1:3}]$ based on cosine similarity and summarize them using a lightweight LLM. Ultimately, we obtain rich object-level multi-source knowledge from the website and offline database.

3.2 Prompt Matching

As for the RAG contents obtained from Parallel Visual Search module, we categorize them into three scenarios and dynamically adjust the corresponding prompts to enhance the model's response quality. Hence, the prompts in the Prompt Templates Pool can be mainly classified into three types.

1) Multi-object multi-source RAG prompt (Appendix A.3): In this case, the image contains multiple objects, while retrieved object-level RAG contents are from multiple RAG sources (e.g., different online search engines and different offline domain-knowledge databases). We designed a prompt suitable for multi-object scenarios and proposed a Visual-RAG alignment module to explicitly align the visual objects with the corresponding text-format object-level RAG contents, enhancing the model's ability to comprehend image information. A detailed illustration of the Visual-RAG alignment module can be found in Section 3.3 and Fig.3.

2) Single-object multi-source RAG prompt (Appendix A.2): When there is only one object in the image

but with multiple RAG sources, we propose a self-rectification mechanism. Instead of assuming that the RAG content is correct and using it as-is, the self-rectification mechanism first relies on the model's pre-trained knowledge to judge the relevance of RAG content and the visual object (prompting LVLMs to identify whether the RAG content is related to the image objects with pre-trained knowledge).

3) General Prompt (Appendix A.1): When no RAG information is available, a general prompt is used to encourage the model to generate an objective image caption with its own knowledge.

3.3 Visual-RAG Alignment Module

In scenarios with multiple objects and multiple RAG sources, traditional RAG strategies are unable to effectively associate the text-format RAG content with the visual objects. This could hinder the effective utilization of the object-level RAG information, affecting the quality of captioning results. We propose a Visual-RAG Alignment Module (VRAM) to bridge the modality gap between textual RAG contents and visual image objects, enabling the model to better understand the image semantics and use the correct RAG content to describe corresponding object. The VRAM, as shown in Fig.3, inspired by the practices of set of marks[Yang *et al.*, 2023], for
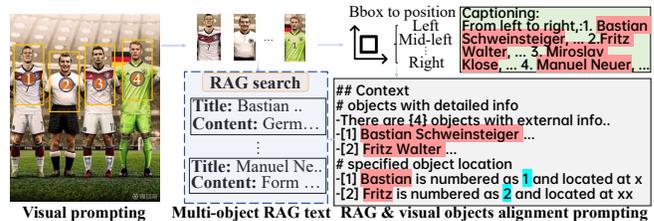

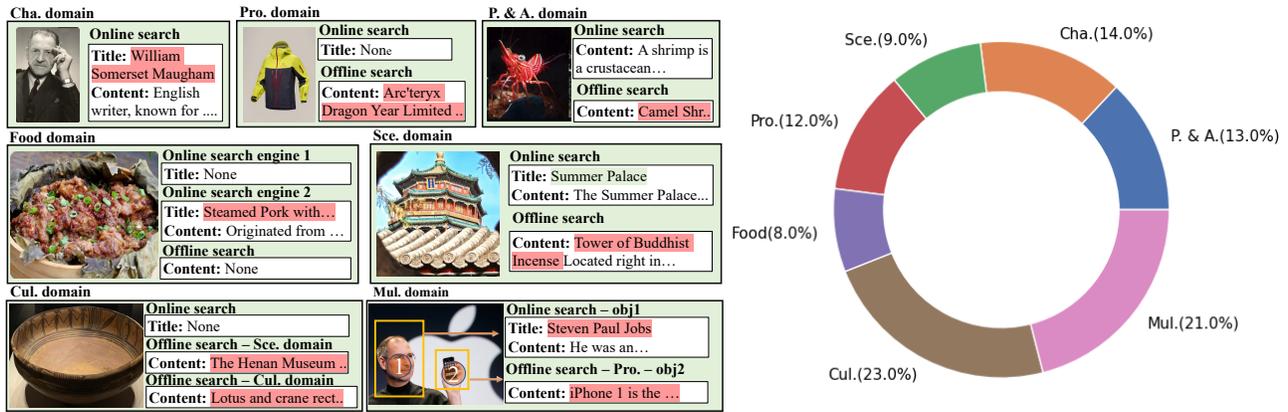Figure 3: Illustration of our VRAM(Visual RAG alignment module).

Figure 4: **Left**: Samples of KAC-dataset in different domains. **Right**: Category statics for KAC-dataset.

input object aware image and object-level RAG contents, we start by marking the objects in the image via their bounding box. Then we formulate visual marks in text format and integrate them with RAG content to construct an object position-aware prompt. Our Visual-RAG alignment module effectively bridges the modality gap between visual and textual modalities. This allows the model to better understand the correspondence between object-level RAG contents and visual objects, thereby outputting high-quality, multi-object knowledge-augmented captions.

4 Experiment

4.1 Settings

Dataset Collection. We classify the open domain images into seven categories, namely Cha.(Characters), A. & P.(Animals & Plants), Sce.(Scenes), Pro.(Products), Food, Cul.(cultural relics), and Mul.(Multi-object scenario). We collect failure cases in real world applications pertinent to our business domain to form the KAC-dataset. The static results of KAC-dataset are shown in Fig.4. The pie chart on the right displays the data proportion for each domain in the KAC-dataset, with Multi-object domain (21%), Culture Relics domain (23%), and Plants & Animals domain (13%) taking up relatively larger shares. The Multi-object domain mainly consists of image data that includes multiple objects, thus it can effectively validate the efficiency of the VRAM within the MsRAG framework. Similarly, the offline domain-knowledge databases within parallel search module aims to address the deficiencies of online visual search engines in retrieving information in areas such as Culture Relics, scenes, and products. So increasing the proportion of data in domains relevant to the offline domain-knowledge databases can verify the efficacy of the parallel visual search module.

In addition to this, we present typical examples from each domain in the KAC-dataset, as shown in the left image of Fig.4. The text in red represents the correct retrieval information. As the online visual search engine performs well in the Cha.(Character) domain, there is no need to implement offline database in Cha. domain. Therefore, data in the Cha. domain typically only contains retrieval contents searched by the online visual search engine. Pro. and P. & A. domains, however, often struggle to retrieve valuable information from the web. The Pro. domain often fails to return any useful retrieval results, while the P. & A. domain usually only retrieves broad classification information about plants or animals, such as their genus or phylum, but lacks detailed species information. For these two domains, the offline domain-knowledge databases always offer more accurate information. It is worth mentioning that since we use more than one online visual search engine, we often get multiple search results from the web, as shown in the example of the Food domain. For the Sce. and Cul. domains, the retrieval information is more complex. In addition to the information retrieved from the web, data in these two domains often simultaneously trigger Sce. and Cul. tracks in the offline database, returning multiple retrieval results. As shown in the example data for the Cul. domain, the retrieval content includes information from both the Sce. track(*Henan Museum*) and the Cul. track(*Lotus and Crane Rectangular Hu*). The *Lotus and Crane Rectangular Hu* is housed in the *Henan Museum*, which means that both pieces of information are meaningful for captioning. Finally, the data from the Mul. domain returns object-level knowledge once it has passed through the Parallel Visual Search Module. Each object's information may stem from either online or offline. As shown in the example from the Mul. domain, object 1 is Steve Jobs, whose information was retrieved online, while the iPhone 1(object 2) he holds was sourced from Pro. track in offline database. Detailed information of offline domain-knowledge database is shown in Table 1, We categorize the data in the offline database into four tracks, including the P. & A. track (Animal and Plant track), the Cul. track (Culture Relics track), the Sce. track (Scenes track), and the Pro. track(Products track). The term 'Entity' represents the number of distinct entities within each track, while 'Volume' denotes the total amount of samples in each track.

Baselines. We evaluate MsRAG on multiple LVLMs, including GPT-4o, Claude-3.5-Sonnet, Qwen2-VL, and InternVL2. For closed-source models (GPT-4o, Claude), we use their APIs; for open-source models, we deploy them with vllm[Kwon *et al.*, 2023]. We also compare MsRAG against mRAG-integrated baselines to verify its effectiveness.

Domain	Image sample	Knowledge sample	Entity	Volume
Animal		Scomber is a genus of ray-finned fish in the family Scombridae living in the open ocean found in Atlantic...	76,683	370,000
Plant		Philodendron tatei 'Congo' is a cultivar of Philodendron tatei of the Araceae family...	152,745	720,000
Culture Relics		The Temple of Dendur (Dendoor in the 19th century) is a Roman Egyptian religious structure originally located in Tuzis...	903,071	2,333,740
Scenes		The statue of Mazu, located at the southern end of Huinu Bay in Chongwu Town, This statue of Guanyin is 32.3 meters high...	5114	34,490
Products		Time for some Tobacco Road vibes with these Jordan 1 Retro Off-Whites. Also known as the "UNC" editions ...	5334	464,305

Table 1: Examples of offline domain-knowledge databases in parallel visual search module.

Evaluations. We evaluate MsRAG on LVLMs using three datasets: CapFusion, Kale, and KAC-dataset. CapFusion and Kale are public captioning datasets with real-world knowledge, aligning well with the knowledge-augmented captioning task, effectively testing MsRAG's retrieval and utilization of external information without queries. All experiments were run on two Nvidia A100s.

4.2 User Study

We conducted a user study on the KAC-dataset using mRAG and MsRAG methods to invoke various LVLMs for captioning the samples in KAC-dataset, with 10 caption results for each sample. Users were asked to rank these captions based on the accuracy and richness of knowledge information they provide. The scoring process was done in a fully blinded manner as the content shown to the users were randomly shuffled results of various captioning methods, ensuring fairness in the scoring process. We eventually collected 20 valid scoring results, and the summary is shown in Table 2.

MsRAG outperformed mRAG in most scenarios, with significant gains in the multi-object (2.78), cultural relics (3.24), and product domains (3.18), demonstrating the effectiveness of its offline database and Visual-RAG alignment. However, the improvement in the character domain was marginal (0.28), as pre-trained LVLMs and online visual search engines already excel in character recognition.

4.3 Quantitative Comparison

Following the evaluation metrics of current image captioning and VQA works, we adopt these metrics to evaluate the performance of our MsRAG in the CapFusion and Kale datasets.

Model	Method	A&P	Cha.	Sce.	Pro.	Cul.	Food	Mul.	Overall
GPT4o	mRAG	6.5	6.4	6.4	5.5	5.4	6.2	6.1	6.1
	MsRAG	**7.9**	**6.6**	**7.5**	**8.3**	**8.5**	**8.7**	**8.0**	**7.9**
Claude-sonnet3.5	mRAG	4.6	4.4	4.9	4.7	4.0	4.9	4.4	4.5
	MsRAG	**7.0**	**6.3**	**6.9**	**7.9**	**7.9**	**6.8**	**7.5**	**7.2**
Qwen2 VL	mRAG	3.3	3.5	2.7	2.0	2.3	2.6	3.2	2.8
	MsRAG	**5.4**	**3.6**	**6.2**	**5.8**	**6.4**	**5.0**	**6.1**	**5.5**
InternVL2	mRAG	3.2	5.4	2.6	2.3	1.7	2.7	2.7	2.9
	MsRAG	**4.8**	**5.5**	**4.8**	**5.5**	**5.8**	**4.9**	**5.5**	**5.3**

Table 2: User study on KAC-dataset.

Evaluation results are displayed in Table 3. It can be inferred that (1) GPT4o combined with our MsRAG has the best overall performance on both the CapFusion and Kale datasets. (2) Our method has continuous gains on almost all of the LVLMs (especially commercial models such as GPT and Claude, which can better exploit multi-source RAG contents). It is noteworthy that metrics like BLEU and CIDEr yielded relatively lower scores on the CapFusion dataset as opposed to the Kale. This is due to the fact that most of the ground truth data in the CapFusion are relatively short and knowledge-oriented, lacking a detailed description of the images. However, generated text often contains more detailed descriptions, which leads to text length significantly surpassing the ground truth text and hence the scores by length-sensitive metrics such as BLEU and CIDEr are skewed.

4.4 Ablation Study

Fig.5 shows our ablation results. We first evaluated VRAM's effectiveness in multi-object scenarios by comparing full MsRAG against baselines without SoM and prompt alignment. Human evaluators scored captions (0=no real-world knowledge, 0.5=partial knowledge, 1=full external knowledge utilization). Results (Fig.5a) show VRAM significantly improves performance (w/VRAM:0.77 vs w/o:0.41), with 86% of scores in the 0.5-1 range (vs 62% without). We also selected two typical examples from the Multi-object domain as shown in Fig.5b. From the above examples, it can be seen that the captioning result with VRAM correctly understands the location of the object in the image, the content of the news and their correspondence with retrieved object-level RAG contents.(eg. Li Yuyi is the official mentioned in the news that is being publicly tried). However, the captioning result without VRAM cannot effectively use retrieval information, confuses the retrieval information and the content of news text in the image (eg. mistakenly regards Liu xiang in the content of news text as the official (Li Yuyi) under public trial). This further proves that VRAM in MsRAG can help LVLM better understand the relationship between the objects in the image and the corresponding retrieved RAG contents.

In addition, we also verified the effect of the offline domain-knowledge databases in the parallel visual search

model	method	CapFusion					Kale				
		B-1	B-2	B-3	M	C	B-1	B-2	B-3	M	C
GPT4o	mRAG	3.3	1.1	0.4	21.4	2.0	27.8	8.3	2.9	14.5	5.4
	MsRAG	**5.6**	**3.4**	**2.1**	**31.3**	**5.2**	**53.2**	**19.6**	**9.4**	**19.9**	**16.7**
Claude-sonnet3.5	mRAG	2.3	0.6	0.5	22.7	1.8	52.2	14.1	4.8	22.6	13.1
	MsRAG	**4.3**	**1.9**	**1.0**	**29.2**	**2.1**	**55.8**	**18.1**	**7.1**	**23.9**	**15.9**
Qwen2 VL	mRAG	2.5	0.7	0.3	22.4	0.3	**62.1**	19.8	7.6	**23.5**	**16.8**
	MsRAG	**5.4**	**2.7**	**1.6**	**30.4**	**1.7**	58.1	**20.5**	**9.5**	23.1	13.3
InternVL2	mRAG	2.3	0.5	0.2	12.1	0.7	28.1	6.6	2.2	14.4	5.4
	MsRAG	**3.7**	**1.2**	**0.6**	**14.5**	**1.8**	**39.2**	**12.6**	**5.6**	**17.3**	**11.1**

Table 3: Quantitative experiments on CapFusion and Kale. B-1 indicates Bleu-1, B-2 is Bleu-2, B-3 is Bleu3, M is METEOR and C is CIDEr.
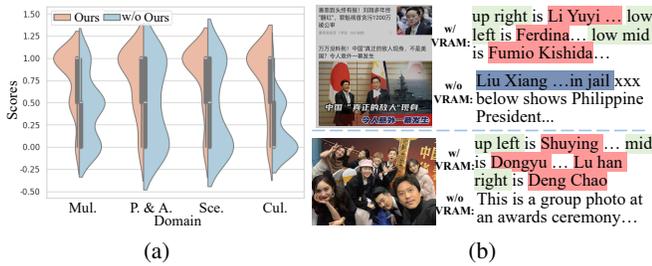

Figure 5: Ablation study. (a) Analysis of the effect of VRAM and offline database. (b) Samples of VRAM ablation experiment

Figure 6: Comparison of MsRAG and m-RAG on VQA settings.

module. We removed the offline database as the baseline. The scoring results are shown in the P. & A. , Sce. and Cul. columns of Fig.5a. After adding the offline domain-knowledge databases, the quality of LVLM-caption in the P. & A. , Sce. and Cul. domain significantly improved (P. & A w/ OD: 0.75, P. & A w/o OD: 0.48. Sce. w/ OD: 0.79, Sce. w/o OD: 0.44. Cul. w/ OD: 0.77, Cul. w/o OD: 0.21.), proving the necessity of using offline database to make up for the online visual search engine and verifying that our prompt design in MsRAG can make LVLM better utilize the complicated multi-source RAG information. Furthermore, We investigated the retrieval information gain of the offline database in the five domains presented in Table 1 on the KAC-dataset. Analysis results are shown in Table 4, the Gain proportion represents the proportion of data where the offline database can provide effective information while online visual search engine fails. It can be observed that the offline database has a continuous gain across all five domains, with the largest gain in the Cul. and Sce. domain.

To further demonstrate MsRAG's superiority over mRAG-based methods in query-free scenarios, we compared them under the VQA setting to highlight key differences. As illustrated in Fig.6, in the first example where the image has multiple objects and the query lacks semantic detail,

MsRAG aligns visual and textual cues effectively, ensuring high-quality responses. Conversely, m-RAG struggles due to insufficient query signals. When the image has a single dominant object and the query is semantically rich (the second example), m-RAG performs better by leveraging precise knowledge retrieval. Overall, MsRAG excels with multi-object images and weak queries, while m-RAG works better with informative queries and single-subject images. Combining both approaches could further enhance VQA performance. Thus, for caption task (like VQA tasks with semantically weak queries), our MsRAG is more effective.

5 Conclusion

In this paper, we propose MsRAG, a general tuning-free framework for knowledge-augmented image captioning. MsRAG is plug-and-play that can be easily, flexibly, and cost-effectively integrated with any open-source or close-source LVLMs. It includes: (1) a Parallel Visual Search Module for retrieving object-level information online and offline without language queries; (2) a Prompt Template Pool to dynamically adjust prompts based on retrieved content, enhancing the LVLMs' ability to utilize complex RAG contents; (3) a Visual RAG Alignment Module to align text-format object-level knowledge with visual objects, enhancing the model's understanding of the correlation between image contents and RAG contents. The aligned image and prompt are fed into the LVLM to generate knowledge-centric captions. Our method is evaluated on the proposed KAC-dataset, CapFusion, and Kale, demonstrating its effectiveness.

Domain	Animal	Plant	Culture Relics	Scene	Product
Gain proportion	37%	31%	62%	52%	43%

Table 4: Analysis of gain from offline database on different domains.

References

[Agrawal *et al.*, 2019] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[Asai *et al.*,] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

[Awadalla *et al.*, 2024] Anas Awadalla, Le Xue, Manli Shu, An Yan, Jun Wang, Senthil Purushwalkam, Sheng Shen, Hannah Lee, Oscar Lo, Jae Sung Park, et al. Blip3-kale: Knowledge augmented large-scale dense captions. *arXiv preprint arXiv:2411.07461*, 2024.

[Bai *et al.*, 2023a] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Bai *et al.*, 2023b] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023.

[Bai *et al.*, 2024] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[Caffagni *et al.*, 2024] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826, 2024.

[Chen *et al.*, 2024] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[Deng *et al.*, 2025] Lianghao Deng, Yuchong Sun, Shizhe Chen, Ning Yang, Yunfeng Wang, and Ruihua Song. Muka: Multimodal knowledge augmented visual information-seeking. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9675–9686, 2025.

[Gao *et al.*, 2023] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[Jiang *et al.*, 2024] Yanbei Jiang, Krista A Ehinger, and Jey Han Lau. Kale: an artwork image captioning system augmented with heterogeneous graph. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7663–7671, 2024.

[Kwon *et al.*, 2023] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

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

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2024] Chuanhao Li, Zhen Li, Chenchen Jing, Shuo Liu, Wenqi Shao, Yuwei Wu, Ping Luo, Yu Qiao, and Kaipeng Zhang. Searchlvlms: A plug-and-play framework for augmenting large vision-language models by searching up-to-date internet knowledge. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[Marino *et al.*, 2019] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[McKinzie *et al.*, 2025] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer, 2025.

[Qiu *et al.*, 2024] Jielin Qiu, Andrea Madotto, Zhaojiang Lin, Paul A Crook, Yifan Xu, Babak Damavandi, Xin Luna Dong, Christos Faloutsos, Lei Li, and Seungwhan Moon. Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 247–266, 2024.

[Rafailov *et al.*, 2024] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[Schuhmann *et al.*, 2022] Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Diy corpora: the www and the translator, 2022.

[Suhr and Artzi, 2019] Alane Suhr and Yoav Artzi. Nlvr2 visual bias analysis. *arXiv preprint arXiv:1909.10411*, 2019.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

[XENOS *et al.*, 2023] ALEXANDROS XENOS, Themos Stafylakis, Ioannis Patras, and Georgios Tzimiropoulos. A simple baseline for knowledge-based visual question answering. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[Yan and Xie, 2024] Yibin Yan and Weidi Xie. Echosight: Advancing visual-language models with wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1538–1551, 2024.

[Yang *et al.*, 2023] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[Yin *et al.*, 2023] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[Yu *et al.*, 2024] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032, 2024.

[Zhang *et al.*, 2024a] Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, et al. mr ^{2} ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *arXiv preprint arXiv:2411.15041*, 2024.

[Zhang *et al.*, 2024b] Zhengxuan Zhang, Yin Wu, Yuyu Luo, and Nan Tang. Mar: Matching-augmented reasoning for enhancing visual-based entity question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, 2024.

[Zhao *et al.*, 2023] Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4736–4756, 2023.

[Zhu *et al.*, 2022] Yongjie Zhu, Chunhui Han, Yuefeng Zhan, Bochen Pang, Zhaoju Li, Hao Sun, Si Li, Boxin Shi, Nan Duan, Weiwei Deng, et al. Adscvlr: Commercial visual-linguistic representation modeling in sponsored search. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 444–452, 2022.