Neurocomputing 676 (2026) 133025

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

ReCoD: Enhancing image description for cross-modal understanding via retrieval and comparison feedback mechanism

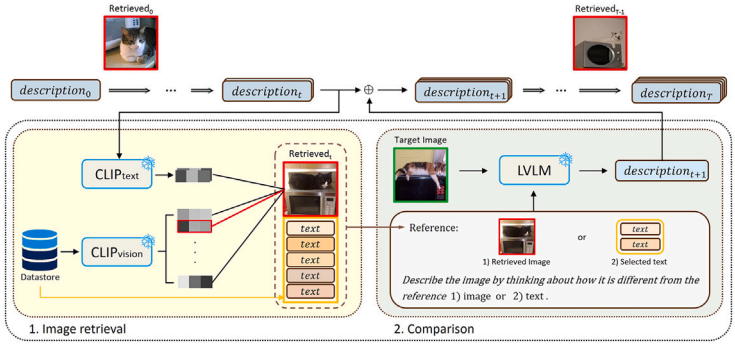
Geunyoung Jung [a], Jun Park [b,1], Hankyeol Lee [a], Kyungwoo Song [c], Jiyoung Jung [a,*]

[a] Department of Artificial Intelligence, University of Seoul, Dongdaemun-Gu, Seoul, 02504, South Korea
[b] Robotics Program, Korea Advanced Institute of Science and Technology, Yuseong-Gu, Daejeon, 34141, South Korea
[c] Department of Statistics and Data Science, Yonsei University, Seodaemun-Gu, Seoul, 03722, South Korea

HIGHLIGHTS

- We propose ReCoD, a training-free method that enhances image descriptions through an iterative feedback mechanism.
- The feedback mechanism consists of two complementary stages: image retrieval and comparison.
- ReCoD is applicable to any LVLM including black-box models, and strengthens vision-language connections, contributing to multi-modal research.

GRAPHICAL ABSTRACT



ARTICLE INFO

Communicated by M. Xu

Keywords:
Vision-language models
Cross-modal understanding
Image-to-text generation

ABSTRACT

To effectively utilize the large language models (LLMs) in the vision domain, it is essential to establish a strong connection between the visual and textual modalities. While deep embeddings can facilitate this connection, representing images as detailed textual descriptions offers significant advantages in terms of the usability and interpretability inherent in natural language. In this paper, we introduce a method of image description enhancement designed to generate highly detailed descriptions that include discriminative attributes of the given image, without requiring additional training. Our method, RECOD, consists of two main components: 1) *"image retrieval"*, which retrieves the image most similar to the descriptions of the target image, and 2) *"comparison"*, which identifies and describes the differences between the target image and the retrieved image. These two components are complementary and form an iterative feedback mechanism. As this process iterates, the retrieved image becomes visually closer to the target image, and the descriptions become progressively more informative. Extensive experiments demonstrate the effectiveness of bridging the gap between the two modalities and the quality of our enhanced descriptions. The code is available at https://github.com/gyjung975/ReCoD.

* Corresponding author.
Email addresses: gyjung975@uos.ac.kr (G. Jung), jun.park@kaist.ac.kr (J. Park), leehk@uos.ac.kr (H. Lee), kyungwoo.song@yonsei.ac.kr (K. Song), jyjung@uos.ac.kr (J. Jung).

[1] Work done at University of Seoul.

https://doi.org/10.1016/j.neucom.2026.133025
Received 25 March 2025; Received in revised form 11 December 2025; Accepted 10 February 2026
Available online 11 February 2026

1. Introduction

The field of multi-modal research is rapidly advancing with vision-language models (VLMs) that integrate vision and language to address a wide range of real-world challenges. Among the pioneering efforts, CLIP [1] has made significant contributions to the field by mapping deep embeddings of visual and textual data into a shared space. However, its encoder-only architecture limits its applicability to specific tasks. Subsequent models have incorporated generative language models [2–4] to improve usability. They can perform various tasks with a generative language models, ranging from image captioning to visual question answering. Recently, the integration of high-performing large language models (LLMs) has notably boosted the performance of VLMs, resulting in large vision-language models (LVLMs). Although they address their dependence on textual decoder efficacy, they continue to lag behind the capabilities of LLMs, primarily due to the challenges in aligning the two modalities.

To date, deep embeddings have been used in all LVLMs to bridge the two modalities. Visual embeddings are mapped to a textual embedding space that LLMs can comprehend, via a lightweight network designed specifically for this purpose. As another way of connecting two modalities, De-Diffusion [5] has shown that precise and comprehensive text can serve as an effective cross-modal interface. Since text is the native input format for LLMs, training for modality alignment is unnecessary. More importantly, regardless of what the task is and which LVLMs is being used, many tasks can be done with any LLM, once the images are represented as text. Therefore, significant advantages can be gained by simply representing images with text.

Image captioning is the most common and simplest way of representing images in text. It involves generating brief and concise captions for a given image, requiring a blend of image understanding and text generation abilities. As a fundamental vision-language task, it has been achieved by the models that follow the architecture of typical VLMs. While effective, these models face scalability issues and struggle to adapt to new, unseen data due to the dependency of visual embeddings on a specific textual decoder, which necessitates retraining whenever the decoder is changed. To overcome these challenges, an alternative strategy is to freeze the visual and textual networks and train only a lightweight network to bridge the two modalities, which has shown substantial success [6,7]. Despite notable advancements, image captions are short, high-level summaries of the images that do not fully represent the visual content.

The emergence of large language models [8–10] has driven the development of more capable VLMs [11–15], commonly referred to as LVLMs. They are not only superior to earlier image captioning models but also able to generate detailed textual descriptions of images. By leveraging vast image-text datasets and extensive modeling, LVLMs create more diverse and detailed descriptions. Despite their success, challenges remain. They often focus primarily on the most prominent features of the main object, overlooking subtler details such as colors, object arrangements, and camera perspectives. Moreover, their ability to produce longer and more descriptive descriptions raises the risk of generating misleading or incorrect information not present in the image, a problem known as hallucination.

In this paper, we propose RECOD, a novel approach to iterative description enhancement aimed at producing detailed and discriminative descriptions for cross-modal understanding. Inspired by the *"Spot the Difference"* and natural human observation skills, RECOD enhances description generation by comparing a target image with a similar one. As illustrated in Fig. 1, by merely providing a similar image, the model's focus shifts to finer details. Our method consists of two components: 1) *"image retrieval"*, which finds the most similar image to the generated descriptions, and 2) *"comparison"*, which identifies and describes the differences between the target image and the retrieved image. There are two types of comparison: image-based and text-based, where the comparison targets are the retrieved image itself and the paired text of the retrieved image, respectively. Through iterative loops, the retrieved image becomes visually closer to the target image, and the generated description correspondingly becomes more detailed and informative about the target image. This process establishes an iterative feedback mechanism between these components, allowing our approach to progressively enrich the detail of the descriptions without the need for training. In addition, RECOD is a plug-and-play module that can be applied to any LVLM, including black-box models accessible only via APIs, without requiring internal model access.

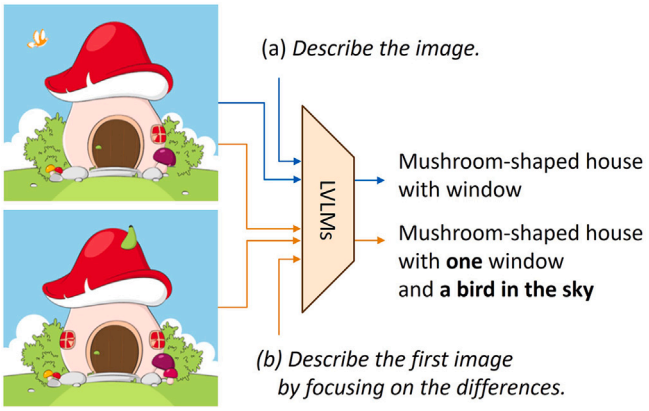


Fig. 1. Illustration of description enhancement by comparison. (a) With a single image and a naive prompt, the LVLMs provides a general scene description. (b) Given a similar comparison target together, they generates a more detailed description by focusing on differences, i.e., "one window" and "a bird".

Extensive experiments in knowledge-based visual question answering (VQA) demonstrate the effectiveness of our method. Given recent attempts to convert knowledge-based VQA to QA to leverage LLMs, we conduct it to demonstrate the value of our descriptions as a cross-modal interface. RECOD yields competitive or even superior performance compared to task-specific training methods. This suggests that RECOD can be utilized as an alternative to deep embeddings for connecting the two modalities. Furthermore, the POPE-based hallucination [16] experiments and qualitative results demonstrate that it effectively reduces hallucinations from two perspectives: the descriptions themselves and the cross-modal interface. Therefore, generating such detailed descriptions with our RECOD goes beyond the simple task of image description generation, and has a broader impact on multi-modal research.

The main contributions are summarized as follows:

- We introduce RECOD, a novel approach that improves image descriptions through an iterative retrieval and comparison feedback mechanism in a training-free manner. It can be integrated into any LVLM, even black-box models.
- By generating detailed and comprehensive descriptions, RECOD strengthens the vision–language connection and mitigates hallucinations from two perspectives, thereby advancing multi-modal research.
- We achieve state-of-the-art performance in knowledge-based visual question answering (VQA) tasks, even including methods that require a training step.

2. Related work

2.1. Vision-language models

Vision-language models (VLMs) [1,17,18] are trained on large-scale image-text pairs using contrastive learning to align visual and textual modalities within a shared embedding space. By using text as labels, they demonstrate strong zero-shot capability and can perform open-vocabulary tasks across a wide range of vision applications, such

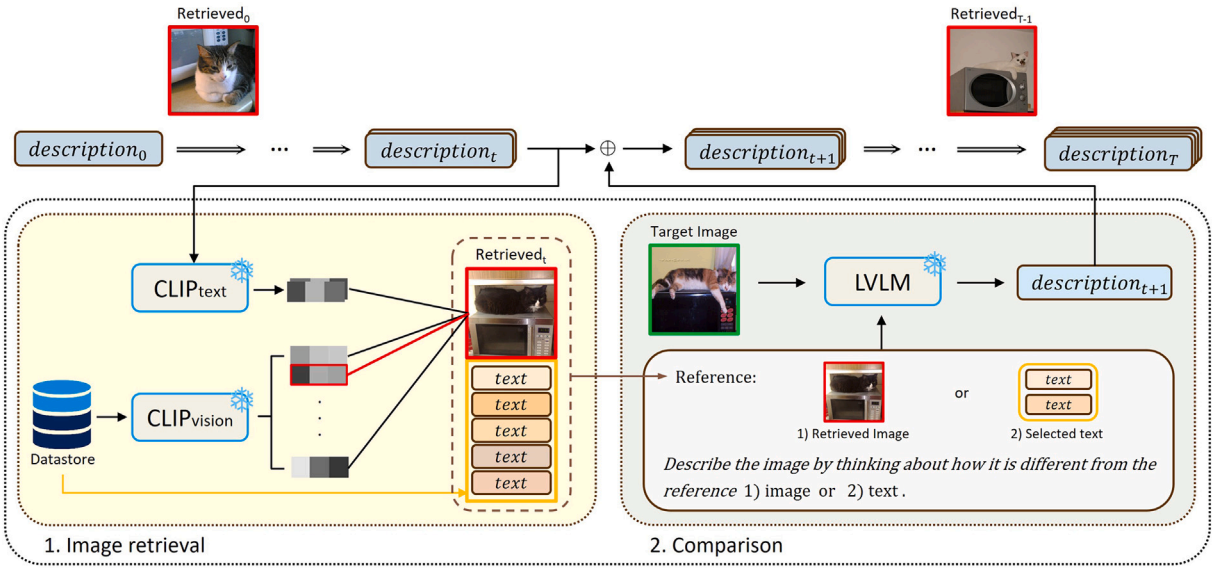


Fig. 2. Overall framework of RECOD. 1) The image retrieval stage retrieves an image using the generated descriptions of the target image based on the similarity of the CLIP's features. 2) Using the retrieved image as a reference, LVLM generates fine-grained descriptions of the target image. The reference, i.e., the comparison target, can be the retrieved image itself or two randomly selected entries from the paired text set for the retrieved image. The retrieved images and the target image are framed in red and green, respectively.

as object detection and semantic segmentation. Building upon VLMs, open-vocabulary object detection methods [19–22] are proposed to recognize novel categories beyond pre-defined label sets, overcoming the limitations of traditional approaches [23–26]. Similarly, VLMs enable open-vocabulary semantic segmentation [27–29], extending their zero-shot capability to dense prediction tasks.

In addition, they also made significant progress in multi-modal tasks, with image captioning being the most representative task. Image captioning is the task of generating short and concise text that provides a summary for an image. Recent advances in this task have focused on scaling the data and model size, substantially increasing the cost of pre-training and fine-tuning [30,31]. Instead of jointly training two networks from both modalities, some previous works like ClipCap [6] and I-Tuning [7] use pre-trained models that remain frozen while only a lightweight connection module is updated. Although these approaches significantly reduce computational resource, they have a major drawback due to the low quality of the training dataset. They are trained to mimic captions that are too short and ignore unique image details. To generate more visually descriptive captions, DiscriTune [32] trains its text decoder via reinforcement learning, using self-image retrieval score of the generated captions as a reward. By optimizing for self-image retrieval rather than replicating human-annotated captions, it mitigates dataset quality issue and succeeds in capturing more discriminative details. More recently, memory-based approaches [33–35] proved their effectiveness. For instance, DeCap [33] represents visual embeddings as a weighted sum of textual embeddings stored in memory, which are then fed into the decoder. SmallCap [34] generates captions conditioned on both an input image and related captions retrieved from an external datastore.

However, captions generated by existing image captioning models remain overly generic and often lack the detail needed to effectively represent the images as a cross-modal interface.

2.2. Large vision-language models and image descriptions

Following the foundational work in the vision-language domain by CLIP [1], several notable large vision-language models (LVLMs) like BLIP-2 [36], instructBLIP [13], and LLaVA [37] have emerged. BLIP-2 enhances image-text alignment through an effective querying transformer. InstructBLIP builds on BLIP-2 by adding the ability to follow complex text instructions. LLaVA connects a visual encoder and an LLM for general-purpose vision-language understanding. It integrates visual and linguistic understanding in a more sophisticated way, enabling generation of detailed image descriptions. However, they still have some limitations. Short descriptions may lack detail by focusing only on major objects, while longer descriptions risk inducing hallucinations.

De-Diffusion [5] introduces a unique vision-language model that combines a text-to-image generator, i.e., Stable Diffusion [38], with visual encoder instead of LLMs. It generates text using a visual encoder and query tokens, which are then used as input prompt for the text-to-image generator. It is fine-tuned based on how closely the generated image matches the original one. Although representing images with text instead of deep embeddings is successful, its reliance on word-level sequences rather than complete sentences restricts interpretability and the depth of conveyed content. Moreover, the high computational cost of text-to-image generation limits its scalability and practicality.

2.3. Knowledge-based visual question answering

Knowledge-based visual question answering (VQA) extends the VQA task by requiring additional external knowledge beyond what is visually presented in the image to answer the question. Early studies have utilized open-source databases such as Wikidata [39] and ConceptNet [40] as sources of external knowledge. With incredible advances in LLMs, recent research [41–45] has leveraged them to achieve remarkable results. However, since LLMs are language-only, they are unable to process images directly. To resolve this issue, PICa [41] first converts the image into a caption using an image captioning model, and then prompt the LLMs with this caption instead of the image. Recognizing the limitations of generic captions in PICa, which miss vital visual details and thus negatively impact performance, subsequent works [42–44] have tried to create question-aware captions. The performance of these methods relies entirely on the quality of the textual representations used in place of the input images.

3. Method

The proposed method, RECOD, is a two-stage iterative framework that enhances image descriptions. It does not require any additional training or human intervention, such as prompt engineering or costly annotations. The overall framework of RECOD is illustrated in Fig. 2.

The image retrieval stage finds an image for comparison, and the subsequent comparison stage generates a comparative description using the retrieved image as a reference. Each component is described in Sections 3.1 and 3.2.

3.1. Image retrieval

Most of the image-to-text generation models in vision-language domain rely on supervised learning to train models to replicate human-generated reference text. Due to the low quality of the ground-truth reference text, they have a significant limitation in that the detail and richness of the generated text is bounded to their training data. However, there are no well-curated datasets for training models to generate more detailed descriptions. In the absence of such high quality datasets, we utilize the result of text-to-image retrieval as guidance to improve the detail of the descriptions in a training-free manner.

In the image retrieval stage, we search for an image to use as a comparison target in the following comparison stage. We retrieve an image that is most similar to the generated descriptions from the datastore $\mathcal{D}$. The datastore consists of either image-only data or image-text pairs, depending on the type of comparison, which will be detailed in Section 3.2:

$$\mathcal{D} = \begin{cases} \{I_i\}_{i=1}^{N}, & \text{for image-based comparison} \\ \{(I_i, T_i)\}_{i=1}^{N}, & \text{for text-based comparison,} \end{cases} \tag{1}$$

where $I_i \in D^I$ and $T_i \in D^T$ denote an image and its paired text, respectively, and N is the size of the datastore. Since text-to-image retrieval requires well-connected multi-modal features between the two modalities, we adopt CLIP [1] textual encoder $\phi(\cdot)$ and visual encoder $\psi(\cdot)$ for feature extraction.

Formally, given the generated descriptions of the target image $\mathbf{C} = [c_0, c_1, \dots, c_t]$ and all images in the D^I, their textual features $\mathbf{w}$ and visual features $\mathbf{z}$ can be obtained as follows:

$$\mathbf{w} = \phi(\mathbf{C}) \in \mathbb{R}^{(t+1)\times d} \tag{2}$$

$$\mathbf{z} = \psi(D^I) \in \mathbb{R}^{N\times d}, \tag{3}$$

where d is the dimension of the output features of CLIP encoder. Then, we calculate the similarity between each of the textual and visual features. To find the image most similar to all the given descriptions, retrieval score of each image r_i is computed by averaging its similarity with each description:

$$r_i = \frac{1}{|\mathbf{C}|} \sum_{j=0}^{t} \text{sim}(w_j, z_i), \tag{4}$$

where $\text{sim}(\cdot)$ denotes the cosine similarity. Finally, the image with the highest score is selected as the output image of the image retrieval stage:

$$I_{i^*} = \text{ImageRetrieval}(\mathbf{C}, \mathcal{D}), \quad \text{where } i^* = \arg\max_i r_i \tag{5}$$

Note that the features of the images in the datastore can be pre-computed for efficiency. We use the MS-COCO [46] train split for the default datastore to avoid data leakage during evaluation. Alternatively, any image dataset with or without paired text, such as Conceptual Captions [47] and WikiWeb2M [48], can be used. An ablation study on the datastore is presented in Section 4.4.2.

3.2. Comparison: spot the difference

While recent LVLMs are superior to traditional image-to-text generation models by a large margin, they rely excessively on the scale of the model and the training dataset. To mitigate this dependency, we apply a simple strategy inspired by the game *"Spot the Difference"*. When we describe an image, our responses often focus on its overall appearance or the primary object alone. However, providing a similar image for comparison invokes consideration of a wider range of visual details and differentiating secondary attributes, such as background, counts, color, size, etc. Similarly, models with strong image recognition and understanding capabilities would pay closer attention to the details of the target image when given a comparison target.

Algorithm 1 ReCoD.

Target image x
Initial description c_0, **t$^{\text{th}}$ description** c_t
textual encoder $\phi(\cdot)$, **visual encoder** $\psi(\cdot)$
Datastore $\mathcal{D} = \{(I_i, T_i) | I_i \in D^I, T_i \in D^T\}_{i=1}^{N}$

```
C = [c_0]
z ← ψ(D^I)
for t = 1, 2, ..., N_loop do
    w ← φ(C)
    r_i ← avg_j(sim(w_j, z_i))
    i* ← arg max_i r_i
    c_t ← LVLM(x, template(RandSample(T_i*)))
    C.append(c_t)
end for
return C
```

In the comparison stage, we prompt the LVLMs to identify the distinguishing visual features of the target image compared to the similar image retrieved in the previous image retrieval stage. There are two types of comparison: image-based and text-based, each utilizing the retrieved image itself and paired text of the retrieved image, respectively. Given the target image x and the result of previous image retrieval stage R, the comparative description c is generated as follows:

$$c = \textbf{LVLM}(x, R), \quad \text{where } R = \begin{cases} I_{i^*}, & \text{for image-based comparison} \\ T_{i^*}, & \text{for text-based comparison} \end{cases} \tag{6}$$

For text-based comparison, when T_{i^*} contains more than two texts, we randomly sample two of them for use. This sampling process corresponds to the "RandSample" operation in Alg. 1. Note that image-based comparison has the significant advantage of utilizing a datastore consisting solely of images, whereas text-based comparison requires a datastore with paired image-text data.

To guide the LVLMs in generating descriptions via comparison, we construct input prompts based on the corresponding instruction templates:

> *<image-based comparison>*
> Reference:
> I_{i^*}
> Describe the first image by thinking about how it is different from the second image, without mentioning the difference, but giving details only about the first image.

> *<text-based comparison>*
> Reference descriptions:
> T_{i^*}
> Describe this image by thinking about how it is different from the descriptions above, without mentioning the difference, but giving details only about the image.

Finally, with these templates, the process of comparative description generation is as follows:

$$c = \textbf{LVLM}(x, \text{template}(R)) \tag{7}$$

In contrast to a naive prompt, e.g., "Describe the image in great detail", prompting LVLMs to describe the image by concentrating on differences between the image and the given comparison target stimulates

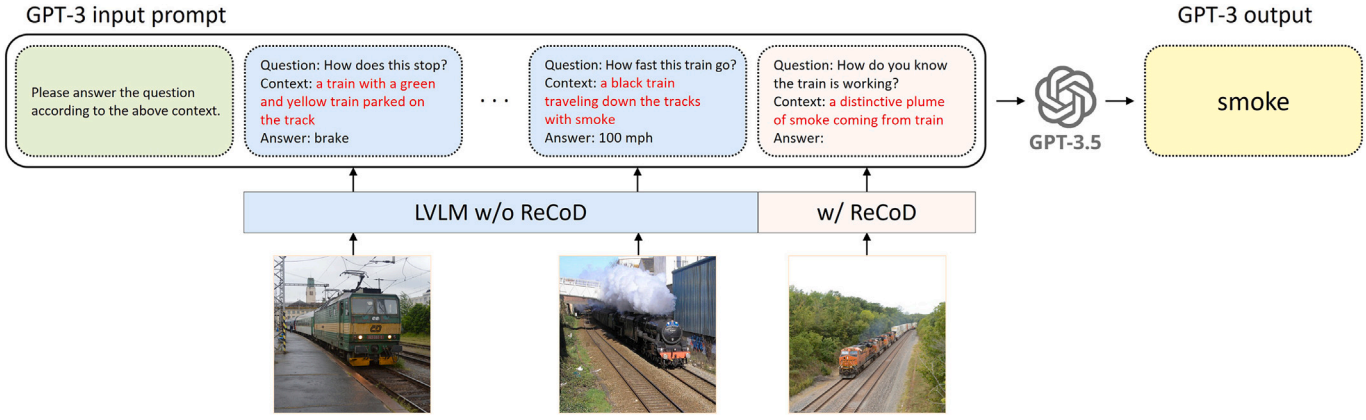


Fig. 3. The pipeline of GPT-3 in-context learning for knowledge-based VQA inference. The red contexts of each question are the descriptions generated by LVLMs. Text-only GPT-3 conducts VQA by replacing the images with descriptions. The input prompt is composed of the user instruction, the in-context examples, and the test sample.

them to capture the specific details. The resulting description is then fed back into the image retrieval stage to retrieve a more similar image, thereby increasing the chance of detecting more subtle differences in the subsequent comparison stage.

3.3. Iterative feedback mechanism

Our method combines the two aforementioned stages—image retrieval and comparison—into an iterative feedback mechanism. The outputs of both stages are complementary and act as feedback for each other. The more similar an image is used for comparison, the finer and more differentiated observations can be made by shifting the model's gaze to subtle differences. Also, detailed descriptions that provide specific visual features help to retrieve a more similar image in the subsequent image retrieval stage. In the context of this mutual feedback setting, in which the retrieved images inform the quality of the generated descriptions, we perform image retrieval based on the generated descriptions rather than on the target images. As a result, this complementary feedback mechanism enables us to gradually enrich the descriptions without careful prompt design, additional training, large datasets, or human supervision.

The overall algorithm of text-based comparison is described in Alg. 1. In case of image-based comparison, it does not require an image-text paired datastore, thus there is no need for D^T. Also, RandSample(T_{i^*}) in line 7 would be replaced with the retrieved image I_{i^*}.

3.4. VQA inference with GPT-3

To evaluate the effectiveness of ReCoD, we conduct a knowledge-based visual question answering (VQA) task. Following the trend of leveraging LLMs for this task by converting images to text, we also adopt GPT-3 [49]. Since performance in this setup relies heavily on how effectively the text represents the original images, this task allows us to clearly assess the impact of ReCoD. We begin by generating descriptions for each image using ReCoD, replacing the images with these descriptions as context. Consistent with prior works, we employ the in-context learning paradigm, where LLMs are guided to perform new tasks by providing relevant examples. Fig. 3 illustrates the inference pipeline using GPT-3.

3.4.1. In-context example selection

As in-context learning becomes increasingly important for effectively using LLMs, research has consistently shown that LLM performance is strongly influenced by the examples provided. Hence, it is essential to select examples that are closely related to the test sample. To ensure relevance, we adopt the approach used in PICa [41], where we calculate the similarity between the test sample and all available examples, and then select the top n examples with the highest similarity scores. Similarity is measured by the sum of the cosine similarities of CLIP embeddings for both images and questions. For consistency, all images in both the examples and the test sample are replaced with their corresponding descriptions.

4. Experiments

4.1. Experimental setting

In this section, we perform knowledge-based VQA experiments to evaluate the quality of descriptions generated by ReCoD. Using the OK-VQA [50] and A-OKVQA [51] datasets, ReCoD consistently improves the performance without training and achieves state-of-the-art results, even outperforming methods specifically designed for this task.

4.1.1. Dataset

We use two knowledge-based VQA datasets, OK-VQA and A-OKVQA. OK-VQA contains 9009 train and 5046 validation image-question pairs. We use soft accuracy from VQAv2 [52] as the evaluation metric. A-OKVQA is the largest knowledge-based VQA dataset and is divided into three splits: 17,056 train, 1145 validation, and 6702 test data. It has a direct-answer (DA) task, which requires correct answers to the open-ended questions, as in OK-VQA, and an additional multiple-choice (MC) task, which selects the correct answer from given choices. The MC task has 4 choices for each question.

4.1.2. Implementation details

We adopt CLIP (ViT-L/14) [1] for image retrieval, and LLaVA-1.5 (13B) [15] and "claude-3-haiku-20240307" of Claude3 [53] for generating comparative descriptions. As mentioned earlier, any LVLM can be used—even a black-box model accessible only via an API, such as Claude3— since ReCoD requires neither training nor internal model access. Unless specified otherwise, LLaVA and Claude refer to LLaVA-1.5 and Claude3, respectively. Since LLaVA is unable to understand two images simultaneously, we only perform a text-based comparison. All models are frozen with no training throughout the entire process.

The maximum output length of the LVLMs and the total number of feedback loops N_{loop} are set to 40 tokens and 10, respectively. We use a *temperature* = 0.2 and a *top_p* = 0.7 as the default parameters for LLaVA with ReCoD. For vanilla LLaVA, both the *temperature* and *top_p* are set to random values for diversity in descriptions. In the case of Claude, the *temperature* is 0.2, while the *top_p* remains at its API default. A single loop means that both stages, image retrieval and comparison, are run one time. All descriptions acquired up to the current loop are used to

Table 1
Results on OK-VQA with the method of representing images, question-awareness, and source of external knowledge. The upper part lists methods that require a training step, while the methods below do not require training and are based on zero-shot or in-context learning. Question-awareness is marked only for methods that use text as image representations.

Method	Image Representation	Question aware	Knowledge-source	Acc(%)
End-to-End Training				
MUTAN [51]	Feature		–	26.4
ConceptBert [54]	Feature		ConceptNet	33.7
KRISP [55]	Feature		Wikipedia+ConceptNet	38.9
MAVEx [56]	Feature		Wikipedia+ConceptNet +Google Images	39.4
KAT [57]	Caption + Tag + Feature	✓	GPT-3 (175B)+Wikipedia	54.4
REVIVE [58]	Caption + Feature	✓	GPT-3 (175B)+Wikipedia	56.6
Prophet [45]	Caption + Answer-candidate	✗	GPT-3 (175B)	57.5
PromptCap [44]	Caption	✓	GPT-3 (175B)	58.4
Zero-shot				
PNP-VQA [59]	Caption	✓	UnifiedQAv2	35.9
LAMOC [42]	Caption	✓	FLAN-T5-XXL	40.3
Img2LLM [43]	Caption	✓	OPT (175B)	45.6
In-context learning				
PICa-Full [41]	Caption	✗	GPT-3 (175B)	46.9
LLaVA-1.5 [15]	Description	✗	GPT-3 (175B)	53.2
+ RECOD	Description	✗	GPT-3 (175B)	55.3
Claude3 [53]	Description	✗	GPT-3 (175B)	52.1
+ RECOD	Description	✗	GPT-3 (175B)	54.4
+ RECOD_{img}	Description	✗	GPT-3 (175B)	53.7

retrieve a similar image, and the images which are retrieved once are excluded from the datastore. The initial description for the first image retrieval is generated by each LVLM without comparison.

4.1.3. In-context learning details

We use "GPT-3.5-turbo-0125" engine for GPT-3 [49] in VQA inference and select the $n = 16$ most similar examples to the test sample in the training split of each dataset. All images of the examples are replaced with a description generated by LVLMs using a naive prompt. Consequently, the only difference between with and without RECOD lies in the descriptions of the test sample.

4.2. Main results

Table 1 compares our RECOD with other methods on the OK-VQA validation set. RECOD and RECOD_{img} represent the text-based and image-based comparisons, respectively. We specify how the images are represented and where the external knowledge comes from. If text is used for image representation, we also mark whether it is conditioned on the corresponding question. The upper part of the table contains methods that require training to perform the task, whether or not they use the training set of OK-VQA for training. The lower part of the table lists zero-shot and in-context learning methods that do not require training. Two baselines, LLaVA-1.5 and Claude3, indicate that the descriptions which replace the test images are generated by themselves with a naive prompt, i.e., "Describe the image in great detail". Each loop for them is entirely independent, with no iterative feedback.

PICa [41] was the first work using GPT-3 as an external knowledge source in knowledge-based VQA. It converts an image into a caption using an image captioning model and then uses it as the context for answering the question. However, it relies on generic captions that are neither relevant to the question nor detailed, and thus does not fully activate the capability of GPT-3. As an alternative approach, later works [42–44,59] generate captions depending on the question. Although they showed promising results, their usability is limited in that the captions are only applicable to specific VQA tasks, i.e., each caption is tied to a single corresponding question. These problems can all be resolved by simply providing highly detailed image descriptions.

Table 2
Comparison with existing methods on A-OKVQA. It has two different types of tasks: direct-answer (DA) and multiple-choice (MC). † denotes training-free methods.

Method	*DA*		*MC*	
	val	test	val	test
ViLBERT [60]	30.6	25.9	49.1	41.5
LXMERT [61]	30.7	25.9	51.4	41.6
ClipCap [51]	30.9	25.9	56.9	51.4
KRISP [55]	33.7	27.1	51.9	42.2
PNP-VQA† [59]	36.0	–	–	–
LAMOC† [42]	37.9	–	–	–
Img2LLM† [43]	42.9	40.7	–	–
GPV-2 [62]	48.6	40.7	60.3	53.7
PromptCap [44]	56.3	59.6	73.2	73.1
Prophet [45]	58.2	55.7	76.4	73.6
LLaVA-1.5†	54.7	51.6	70.4	71.1
+ RECOD†	58.1	53.0	71.4	72.1
Claude3†	50.8	48.5	70.6	71.6
+ RECOD†	52.5	49.9	71.3	72.5
+ RECOD_{img}†	55.8	50.2	72.1	72.7

Even though RECOD is not a specialized method for the knowledge-based VQA task, but only for enhancing image descriptions, it achieves the highest performance (55.3%) among the training-free methods in the lower part of the table. It also shows the remarkable results on A-OKVQA in both direct-answer (58.1%) and multiple-choice (72.7%), as shown in Table 2. These are the best performance, excluding methods that require any training. Additionally, regardless of the comparison type, both types of comparison lead to performance improvements. Since the only difference with baselines, LLaVA-1.5 and Claude3, is the descriptions of the test sample, it is apparent that the improvements have been driven by enhanced descriptions that are sufficiently representative of the image. This suggests that the GPT-3 understands the image

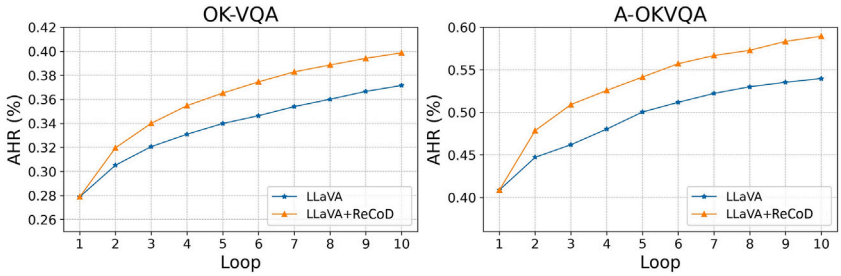


Fig. 4. Comparison of answer hit rate (AHR) over the loops on both datasets. ReCoD is increasing more steeply than LLaVA.

Table 3
Mean and standard deviation of performance. The deviations are significantly reduced when ReCoD is applied.

Method	OK-VQA	A-OKVQA
LLaVA-1.5	52.62 ± 0.42	54.60 ± 0.47
+ ReCoD	**55.35** ± 0.12	**58.16** ± 0.03
Claude3	51.89 ± 0.15	51.12 ± 0.23
+ ReCoD	**54.38** ± 0.09	52.74 ± 0.17
+ ReCoD$_{img}$	53.73 ± 0.08	**55.83** ± 0.07

Table 4
Performance across loops on each dataset. The performance gains are much larger when ReCoD is applied.

Method	OK-VQA			A-OKVQA		
	1	5	10	1	5	10
LLaVA-1.5	52.1	52.7	53.2	51.4	54.0	54.7
+ ReCoD		54.1	**55.3**		55.4	**58.1**
Claude3	49.8	51.9	52.1	48.3	49.5	50.8
+ ReCoD		52.9	**54.4**		51.4	52.5
+ ReCoD$_{img}$		52.5	53.7		54.4	**55.8**

well enough by looking at our descriptions instead of the image itself. Moreover, descriptions from ReCoD are question-independent, completely removing usability constraint. In this sense, we would like to emphasize that the outstanding ability of ReCoD to generate detailed descriptions of the images can be useful in a wide range of multi-modal tasks with LLMs, and can therefore play a valuable role in bridging the two modalities.

4.2.1. Statistical analysis

Table 3 shows the mean and standard deviation of performance across three runs on both datasets. The mean accuracies are consistent with those reported in Tables 1 and 2, and the deviations are significantly reduced when applying ReCoD. The baselines have higher deviations due to less informative descriptions, leading GPT to rely more heavily on its internal knowledge to answer the questions. In contrast, the detailed descriptions generated by ReCoD result in much lower deviations, as they provide sufficient information.

4.3. Feedback mechanism

We conduct extensive experiments to quantitatively measure the effect of the feedback mechanism, which is the key strategy of ReCoD. We use a validation split of OK-VQA and A-OKVQA.

4.3.1. Performance by loop

Table 4 shows the change in accuracy for each loop. We accumulate the generated description at each loop and use all these descriptions as replacements for each image. Loop 1 denotes the use of a single initial description generated by each LVLM via a naive prompt.

Table 5
Results of using LLaMA-3 instead of GPT-3. It achieves state-of-the-art performance while being much smaller than GPT-3.

Method	OK-VQA			A-OKVQA		
	1	5	10	1	5	10
LLaVA-1.5	59.3	59.7	59.9	57.2	58.3	59.3
+ ReCoD		60.7	**61.6**		59.8	**61.2**
Claude3	58.8	60.0	60.0	56.6	56.5	57.4
+ ReCoD		60.8	61.5		58.5	59.2
+ ReCoD$_{img}$		60.9	**61.9**		59.3	**60.2**

On the OK-VQA dataset, both LVLMs with ReCoD consistently show a larger increase in accuracy as the loop progresses compared to without ReCoD. The results are also consistent on the A-OKVQA dataset, and the improvement in ReCoD$_{img}$ is remarkable. The fact that ReCoD increases accuracy much more than LVLMs alone as the loop continues suggests that the performance gain from ReCoD is not merely a consequence of increasing the number of descriptions in the test sample.

4.3.2. Answer hit rate

In this section, we evaluate how meaningful the visual context encoded in the descriptions is for answering the corresponding question by calculating the answer hit rate (AHR). The AHR is the proportion of question-answer pairs that include the ground-truth answer in their context descriptions. As our main purpose is enhancing the image descriptions, we only consider the context descriptions of the test sample, excluding those of in-context examples.

Fig. 4 plots the change in AHR of LLaVA with and without ReCoD over the loops on both datasets. Although both increase consistently, ReCoD has a steeper rise compared to LLaVA. This implies that the feedback mechanism is considerably effective in capturing diverse and detailed information. High AHR naturally leads to high accuracy because it means that more of the strongest cues to answer the question, i.e., exact answer word, are provided. It is quite notable that ReCoD achieves high AHR even though it generates general-purpose descriptions rather than question-conditioned descriptions. In addition, the results further support the reduction in deviations observed in Section 4.2.1.

4.4. Ablation studies

We conduct ablation studies on the inference LLM, target LVLM, datastore, and sampling hyperparameters to understand the impact of each component.

4.4.1. Analysis of LLM impact on performance

To verify whether the performance gains are due to the remarkable capability and huge size of GPT-3 (175B), we conduct knowledge-based VQA under the same settings, only replacing GPT-3 with a smaller LLaMA-3 (70B) [9]. Table 5 confirms that LLaMA-3 also consistently performs better when ReCoD is applied. Surprisingly, ReCoD achieves

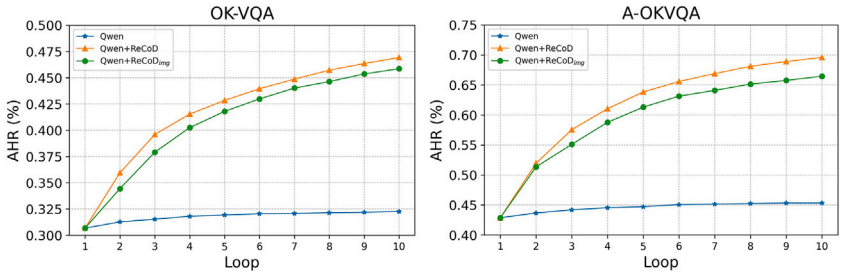


Fig. 5. Answer hit rate (AHR) over the loops on both datasets using Qwen2-VL. Qwen2-VL alone gives a marginal gain, but the increases with ReCoD are remarkable.

Table 6
Results with Qwen2-VL-7B. The effect of ReCoD is strong even for model of size 7B.

Method	OK-VQA	A-OKVQA
Qwen2-VL	52.83	55.41
+ ReCoD	**56.16**	**59.94**
+ ReCoD$_{img}$	54.50	59.65

Table 7
Ablation on two additional datastores. ReCoD leads to significant performance improvements regardless of the size and quality of the datastore.

Method	OK-VQA		A-OKVQA	
	CC3M	WikiWeb2M	CC3M	WikiWeb2M
LLaVA-1.5	53.19	53.19	54.70	54.70
+ ReCoD	**54.24**	**54.20**	**58.64**	**58.57**
Claude3	52.07	52.07	50.82	50.82
+ ReCoD	54.49	**54.05**	52.68	52.56
+ ReCoD$_{img}$	**54.51**	53.84	**56.40**	**56.19**

state-of-the-art performance on each dataset (61.9% and 61.2%), outperforming all existing methods, even those that require a training step. It's worth noting that the same results can be reproduced with the smaller open-source model, as opposed to the GPT-3 API which has a non-negligible cost. Furthermore, the results also demonstrate the usability of ReCoD for any LLM, enabling the effective connection between vision and language.

4.4.2. Ablation on target LVLM and datastore

We performed further experiments with Qwen2-VL-7B [63], which is much smaller than LLaVA and Claude. Table 6 shows that even a 7B model achieves substantial performance gains with ReCoD in both text-based and image-based comparisons. Notably, as illustrated in Fig. 5, Qwen2-VL-7B alone yields minimal improvement in AHR over the loops, whereas ReCoD leads to a significant increase. The successful application of ReCoD not only to black-box models but also to small-scale models demonstrates its broad applicability.

Table 7 presents results using two additional datastores: Conceptual Captions 3M (CC3M) [47] and WikiWeb2M [48], which contain approximately 3.3M and 2M images, respectively. Although both datastores have one text per image and text quality is relatively low, they still yield notable performance improvements. Moreover, the strong performance with the default datastore, COCO—despite being only 5% of their size—demonstrates robustness of ReCoD.

4.4.3. Ablation on LVLM sampling hyperparameters

We ablate the decoding hyperparameters of LVLMs—*max_new_tokens*, *temperature*, and *top_p*—using LLaVA on the A-OKVQA dataset. Table 8 presents the performance with varying maximum output length of

Table 8
Results under different *max_new_tokens* settings.

Method	*max_new_tokens*		
	40	80	120
LLaVA-1.5	54.7	54.4	54.0
+ ReCoD	**58.1**	57.8	57.6

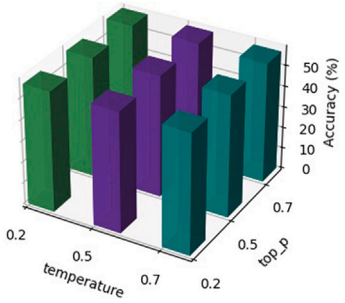


Fig. 6. Sensitivity analysis of *temperature* and *top_p*.

Table 9
Comparison of time and memory for the two stages per loop. Retrieval is a shared part regardless of LVLMs.

Method	Time (s)	Memory (GB)
	Retrieval	
CLIP	0.0085	0.0035
	Generation	
LLaVA-1.5	2.6642	1.1598
+ ReCoD	2.6836	1.1797
Qwen2-VL	1.5219	0.3446
+ ReCoD	1.5360	0.3595
+ ReCoD$_{img}$	1.5837	0.4911

LVLM, i.e., *max_new_tokens*. ReCoD improves accuracy across all settings, with the best result at 40 tokens, which we use by default. This suggests that increasing the output length does not necessarily lead to richer or more informative content. Fig. 6 illustrates the sensitivity analysis of *temperature* and *top_p*. ReCoD remains stable across ranges, indicating low sensitivity to these hyperparameters.

4.5. Time and memory analysis

To investigate the additional cost of applying ReCoD, Table 9 reports the average wall-clock time (s) and GPU memory usage (GB) per feedback loop. We measure the retrieval and generation stages separately. The retrieval stage—which exists only in ReCoD—is shared part independent of LVLMs. As Claude is a black-box model accessible only via

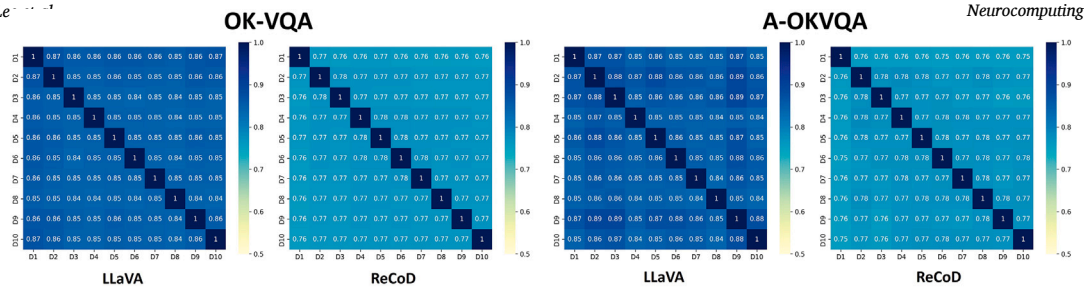


Fig. 7. Heatmaps of cosine similarity of the CLIP textual embeddings between descriptions across loops. The results are the average for all samples in each dataset. The descriptions from ReCoD have low similarity.

Table 10
Results of five reference-based metrics.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
LLaVA	12.18	**26.48**	36.79	1.94	22.97
+ReCoD	**12.84**	26.46	**37.95**	**10.40**	**23.10**

Table 11
Winning rates (%) evaluated by GPT-4o and human. The descriptions generated by ReCoD are always better than baselines.

Method	OK-VQA		A-OKVQA
	GPT-4o	Human	GPT-4o
LLaVA w/ ReCoD	79.3	55.0	79.0
Claude w/ ReCoD	63.7	60.0	54.0
Claude w/ ReCoD$_{img}$	63.0	64.0	71.7

Table 12
Accuracy on the VQAv2 dataset.

Method	VQAv2
LLaVA-1.5	57.9
ReCoD	**60.0**

API, Qwen2-VL is used for the image-based comparison. All experiments are conducted on two NVIDIA A6000 GPUs.

As shown in the top row, CLIP-based retrieval is extremely lightweight, taking only 8.5 ms and 3.5 MB—essentially negligible. The lower block compares generation with and without ReCoD. As described in Section 3.2 about instructions, text-based comparison only adds two reference texts to the text prompt that will be input to the LLMs or LVLMs. Therefore, incorporating ReCoD to LLaVA increases time by just 0.019 s (≈ 0.7%) and memory by 0.02 GB (≈ 1.7%). For Qwen2-VL, the overhead is only 0.014 s and 0.015 GB, underscoring the efficiency of our method.

The image-based comparison ReCoD$_{img}$ introduces slightly larger cost since it processes both the retrieved and target images, resulting in an additional 0.062 s (≈ 4%) and 0.15 GB (≈ 43%). This overhead can be reduced through efficient visual encoding—for example, by caching target image tokens, or by reducing the number of visual tokens, which has recently been actively studied as a major computational bottleneck of LVLMs [64,65].

4.6. Textual quality evaluation

We analyze the quality of our descriptions from the perspective of the text itself. First, we measure five traditional reference-based metrics: BLEU-4 [66], METEOR [67], ROUGE-L [68], CIDEr [69], and SPICE [70]. Table 10 shows that ReCoD achieves comparable or superior results across all metrics. While traditional image-to-text generation models aim to imitate reference text, LVLMs generate longer and more diverse descriptions, making these metrics unsuitable for evaluating open-ended outputs. As an alternative, we evaluate both datasets using GPT-4o [71] and human judgments on OK-VQA. Both GPT-4o and human were asked to choose the better of the two descriptions generated by LVLMs with and without ReCoD. We randomly sample 100 images from each dataset. Table 11 shows the winning rates against the corresponding baselines, i.e., LVLMs without ReCoD. Our descriptions significantly outperform baselines. For GPT-4o, results are averaged over three runs, and for human evaluation, we apply majority voting from three workers.

4.6.1. Diversity of descriptions

In this section, we analyze how diverse the generated descriptions are across feedback loops. Fig. 7 presents heatmaps of the average cosine similarity between the CLIP textual embeddings of descriptions at each loop. Since the baseline LLaVA generates descriptions solely based on the target image, its outputs always remain highly similar. In contrast, ReCoD generates more diverse and semantically varied descriptions, as each loop employs a different comparison target.

4.7. Qualitative results

The five descriptions generated by LLaVA with and without ReCoD are presented in Fig. 8, as well as predicted answers. GPT-3 has to correct the open-ended answer of each question based on these descriptions without the image. It produces the correct answers when using the descriptions of ReCoD as the context of the image, but is unable to do so with the generic descriptions from LLaVA. The exact answer words only contained in the descriptions of ReCoD, e.g., “pelican” or “screen saver”, allow GPT-3 to properly answer the questions. The underlined parts mark additional information that is not available without ReCoD and can be clues to answer arbitrary questions. Keep in mind that all this information comes from general-purpose descriptions that are not conditioned on any specific question. More qualitative results are provided in Appendix B.

Fig. 9 illustrates the patch-wise similarity between the target image and the generated descriptions. The left image shows the similarity to the initial description, while the right image presents the average similarity over all descriptions up to loop 5. The text labeled “Initial” denotes the initial description, and the bullet-pointed texts below correspond to partial descriptions at loop 5. As the loop progresses, the number of patches with high similarity increases and additional fine-grained details emerge, which are highlighted in green.

4.8. Factual information preservation

It is important to generate detailed descriptions while maintaining the factual information of the image. To demonstrate that ReCoD generates rich and informative descriptions while preserving factual

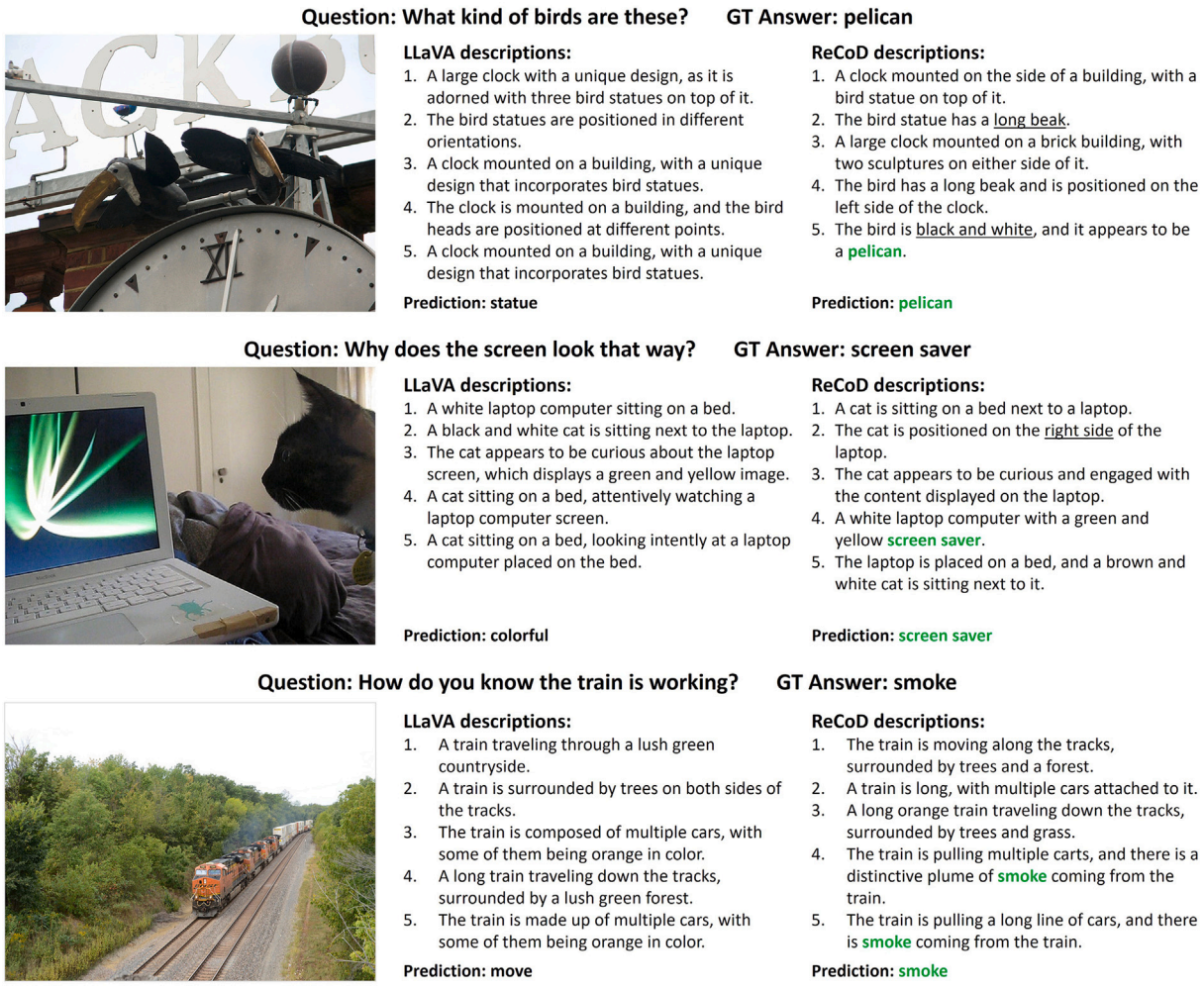


Fig. 8. Qualitative results of generated descriptions on OK-VQA. The ReCoD make the GPT-3 answer correctly by including the ground-truth answer in the descriptions while failing with generic descriptions from LLaVA. We underlined details that are seen only in ReCoD.

content, we conduct the same experiment on the VQAv2 [52] dataset. In contrast to knowledge-based VQA, VQAv2 only contains questions about objective facts visible in the image. Table 12 reports the performance on 5000 samples at loop 10, showing an improvement of 2.1 percentage points when ReCoD is applied. Additionally, Fig. 10 provides qualitative results, where the generated descriptions accurately reflect the factual content of the image, leading to correct answers to the questions.

4.9. Hallucination analysis

4.9.1. Perspective of descriptions

To evaluate the hallucinations in the descriptions, we adopt POPE [16]. POPE is a polling-based approach originally designed for evaluating object hallucinations in LVLMs. Given a ground-truth object list of the image, it queries the LVLMs to answer yes or no regarding whether each object is in the image or not. The hallucinations are evaluated by posing questions about objects that are not in the ground-truth object list and checking whether they answer "no".

However, since we aim to evaluate hallucinations in the generated descriptions rather than in the LVLMs themselves, we modify POPE's pipeline accordingly. Specifically, we first extract object nouns from each description using GPT-4.1 [10]. Because this step is a simple noun-extraction task, the use of GPT-4.1 does not introduce errors. We then follow the POPE procedure by querying whether each extracted object is present in the image. The answers are also obtained via GPT-4.1. As our focus is on hallucinations in the descriptions, we omit the object

Table 13
Hallucination accuracy (%) of generated descriptions across loops on OK-VQA and A-OKVQA, evaluated using the POPE-based pipeline. Accuracies indicate the percentage of "yes" answers. The values in parentheses represent the average number of extracted objects per image.

Method	OK-VQA		A-OKVQA	
	5	10	5	10
LLaVA-1.5	87.6 (4.3)	**86.0** (4.6)	88.6 (4.3)	87.6 (4.6)
+ ReCoD	87.6 (5.0)	85.9 (5.4)	**88.9** (5.1)	**87.7** (5.4)
Claude3	88.8 (5.1)	87.8 (5.6)	90.6 (4.3)	89.5 (5.3)
+ ReCoD	88.9 (5.2)	**88.5** (5.6)	**90.8** (5.3)	**90.0** (5.8)
+ ReCoD$_{img}$	**89.1** (5.9)	88.2 (6.5)	90.7 (6.0)	**90.0** (6.6)

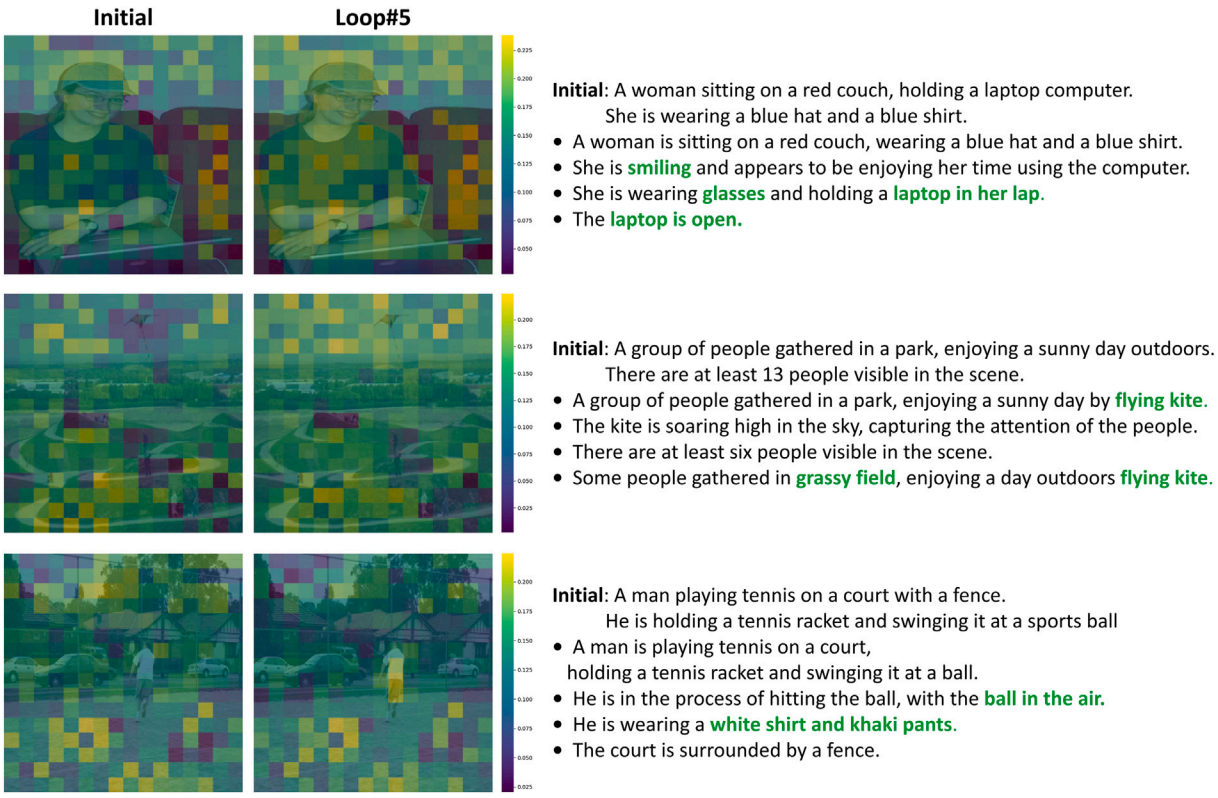


Fig. 9. Patch-wise similarity between the target image and the generated descriptions. The text marked "Initial" represents the initial description, and the bullet points indicate partial descriptions at loop 5. As the loop progresses, patches with high similarity increase, and finer-grained details are provided.

disruptor and treat any object that is identified as not present ("no" answer) as hallucinated.

Table 13 reports hallucination accuracy (%) at 5 and 10 loops, measured as the proportion of "yes" answers, where higher values indicate fewer hallucinations. The values in parenthetical denote the average number of extracted objects per image. Across both OK-VQA and A-OKVQA, RECOD consistently improves accuracy. Notably, it reduces hallucinations while generating more detailed descriptions. Remarkably, the average number of extracted objects per image is even higher, demonstrating that RECOD enhances descriptive richness without sacrificing factual accuracy.

Fig. 11 provides qualitative examples comparing LLaVA and RECOD. Each pair shows a target image, a retrieved image, and the corresponding object lists extracted from the generated descriptions. While LLaVA often mentions only a few dominant objects, RECOD yields richer and more diverse lists by leveraging the retrieved images as comparison targets, highlighting its ability to produce more informative descriptions.

4.9.2. *Perspective of cross-modal interface*

RECOD also has huge advantage for hallucination in the perspective of a cross-modal interface. It can reduce hallucinations in LLMs in situations where they are given the generated descriptions instead of the deep embeddings. Generally, LVLMs only describe what is visible in a given image. This can be potentially problematic when used as a cross-modal interface. Any information that is not explicitly described in the descriptions may result in hallucinated outputs driven by LLMs' internal knowledge or bias, such as co-occurrence or object counts.

Fig. 12 displays the target image and generated description, as well as the retrieved image that is used to generate the description. The bold parts demonstrate that RECOD is able to explicitly negate non-existent objects through comparison. For example, it can identify that there are no clouds when compared to an image of a cloudy sky (top left), or that there is only one plane when compared to an image of two planes (bottom right). Sky and clouds, roads and cars, refrigerators and food, and so on, are usually paired together. Therefore, explicitly informing LLMs of an absence helps reduce hallucinations.

5. Limitations and future works

Although the feedback loop continually enriches the descriptions, it suffers from three limitations: (i) a naive retrieval method, (ii) information redundancy, and (iii) lack of factual content. Leveraging the natural strength of CLIP—which aligns image and text modalities in a shared embedding space—and following its wide adoption in multimodal zero-shot classification [1] and retrieval [32,34], our simple CLIP-based retrieval works reasonably well. However, due to the remaining modality gap [72], this naive retrieval process may miss subtle or fine-grained details. A more advanced retrieval strategy could further enhance the overall performance of the framework. In addition, only small amounts of discriminative detail are added in each loop, while the

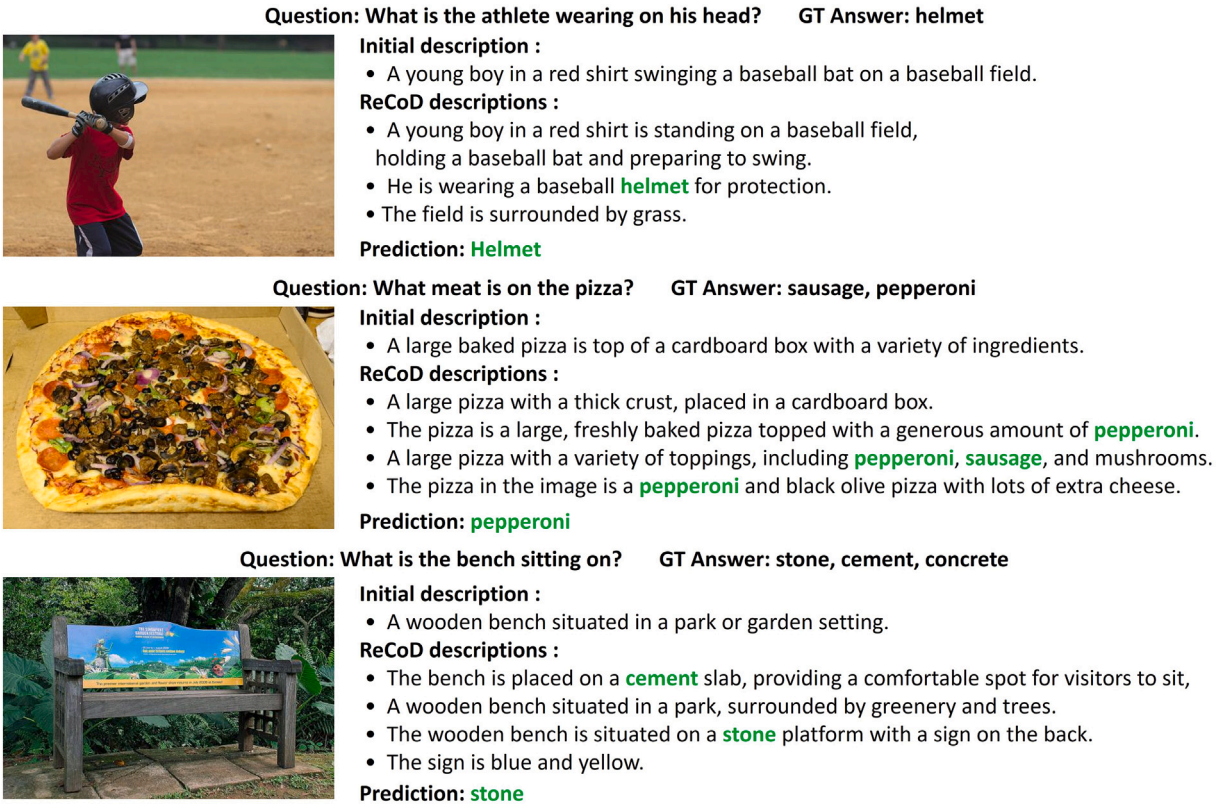


Fig. 10. Qualitative results of generated descriptions on VQAv2. They include factual information about the image, and answer questions correctly. The answer words are highlighted in green.

general base content tends to be repeated, resulting in redundancy in the final descriptions. We attempt to mitigate this by summarizing the final descriptions using LLMs, but distinctive details that appear only once are often omitted during summarization. This redundancy can be effectively addressed either by filtering out generic content during generation or by summarizing in a way that preserves important information. Lastly, since the descriptions are generated through an iterative retrieval–comparison process, they can be influenced by retrieval noise and may emphasize inter-image differences, thereby overlooking essential factual content.

Beyond these limitations, extending ReCoD to the video domain represents another promising research direction. Directly processing entire videos exacerbates redundancy and incurs substantial computational costs, especially for video-based comparisons. Effective video handling further requires explicit modeling of temporal relations, actions, and motion, as well as LVLMs capable of understanding and comparing two videos simultaneously. These challenges are non-trivial and beyond the scope of our current work, but we plan to explore them in future studies.

6. Conclusion

We presented ReCoD, a novel training-free approach for enhancing image descriptions without the need for large-scale datasets. Our framework enables models to iteratively refine descriptions through a feedback loop between retrieval and comparison stages, capturing more fine-grained details. Such highly detailed descriptions serve as a cross-modal interface, effectively connecting the vision and language modalities. As an alternative to deep visual embeddings, it offers significant advantages such as improved usability and interpretability. Furthermore, ReCoD can be easily applied to any LVLM, including black-box models. We demonstrated the quality of the descriptions generated by ReCoD on a knowledge-based visual question answering task, achieving state-of-the-art performance. Additional hallucination analysis reveals that hallucinations can be mitigated from two perspectives: the descriptions themselves and the cross-modal interface. ReCoD contributes to multi-modal research by providing rich and informative descriptions that enhance understanding and interaction between vision and language modalities.

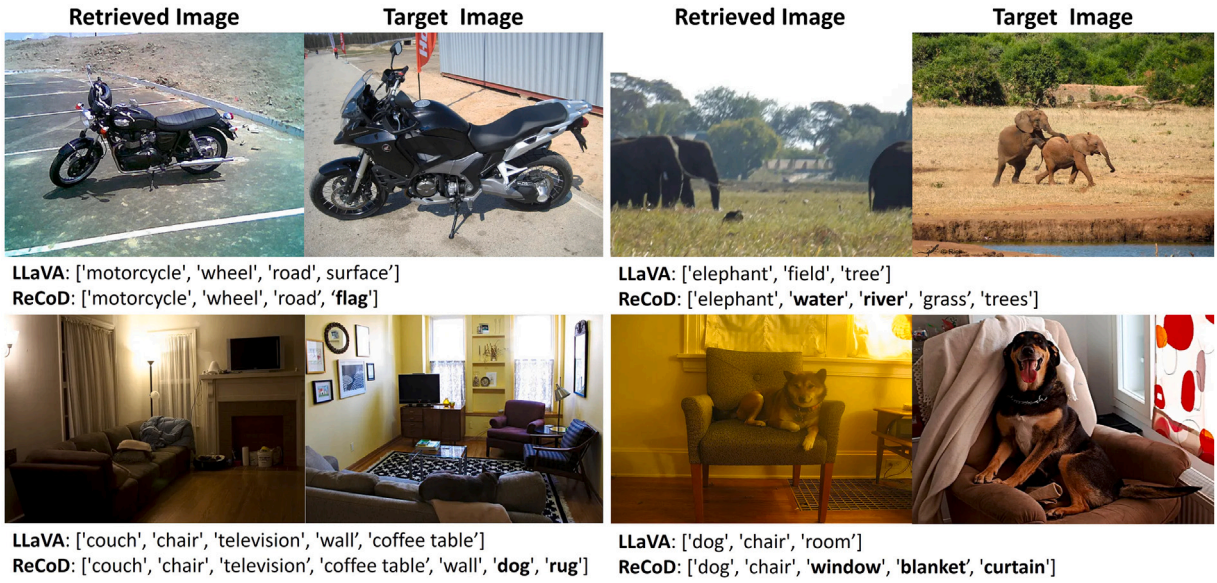


Fig. 11. Examples of a target image, one of the retrieved images, and extracted object lists from the generated descriptions by each method. The list from ReCoD includes more diverse and detailed objects.

Fig. 12. Qualitative analysis of hallucinations from the perspective of a cross-modal interface. The descriptions include the information about "no" as a result of the comparison with the retrieved image, e.g., "no cloud" and "no cars". The bold and underline are object negation and detail information, respectively.

CRediT authorship contribution statement

Geunyoung Jung: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Jun Park**: Writing – review & editing, Writing – original draft, Validation, Software. **Hankyeol Lee**: Writing – review & editing, Writing – original draft, Validation, Software. **Kyungwoo Song**: Writing – review & editing, Writing – original draft, Supervision. **Jiyoung Jung**: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships that may be considered potential competing interests:

Jiyoung Jung has patent pending to University of Seoul. Geunyoung Jung has patent pending to University of Seoul. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the 2024 Advanced Facility Fund of the University of Seoul for Jiyoung Jung. Additionally, this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) for Geunyoung Jung (RS-2025-24523036).

Appendix A. More implementation details

To avoid data leakage, we used COCO [46] 2014 and COCO 2017 train split as the datastore for OK-VQA [50] and A-OKVQA [51], respectively. Each dataset contains 82,783 and 118,287 images with 5–6 paired texts per image. For the initial description and the output of the comparison stage, we removed the non-informative phrases, e.g., "The image shows", in post-processing.

Appendix B. More qualitative results

B.1. Textual quality

Fig. B.13 displays additional qualitative results of descriptions generated by RECOD. The descriptions contain details and supplementary information that go beyond the primary objects, explicitly providing the correct answer to the given question.

Question: What green food is on top of the pizza? GT Answer: spinach

ReCoCap captions:
1. A pizza with a variety of toppings, including tomatoes, cheese, and greens.
2. The pizza is placed on a plate, which is sitting on a dining.
3. The pizza in the image isa personal size pizza with a variety of toppings.
4. The pizza is placed on a paper plate, which is sitting on a wooden table
5. The pizza is topped with a variety of vegetables, including tomatoes, **spinach**.

Prediction: spinach

Question: What kind of carrying equipment is on the floor at the foot of the bed? GT Answer: backpack

ReCoCap captions:
1. A bedroom with a large bed, a couch, and a chair.
2. The bed is positioned near a window, and there are two windows in the room.
3. A bedroom with a large bed, a **backpack**, and a chair.
4. The bed is covered with a blue comforter, and there is a chair nearby.
5. The bed is positioned next to a window, allowing natural light to enter the room.

Prediction: backpack

Question: What type of train is this? GT Answer: passenger

ReCoCap captions:
1. A train on a track, with the train's shadow visible on the ground.
2. The train is traveling through a countryside setting, and there are a few people.
3. The train is a **passenger** train, as indicated by the presence of windows on the side of train.
4. A train on the tracks, with the sun shining on the side of the train.
5. The train is positioned near a platform, and the sun is shining on the tracks.

Prediction: passenger

Question: What is the style of brick on the wall? GT Answer: stone

ReCoCap captions:
1. A small bathroom with a toilet and a sink.
2. The toilet is positioned next to a **stone** wall, and the sink is located on the left side.
3. The toilet is positioned next to a **stone** wall, and the sink is located on the opposite side.
4. The bathroom is small and has **stone** wall.
5. The sink is located to the left of the toilet.

Prediction: stone

Question: Name a metal shown? GT Answer: spoon

ReCoCap captions:
1. A wooden cutting board with a ham sliced into pieces.
2. The ham is accompanied by a variety of vegetables.
3. There is a large piece of meat, possible ham and a **spoon** resting on cutting board.
4. The ham is sliced and arranged on the platter.
5. The food is arranged in a visually appealing manner.

Prediction: spoon

Question: What type of instrument is this? GT Answer: guitar

ReCoCap captions:
1. A group of people sitting around a dining table, enjoying a meal together.
2. There are various food items on the table, including a donut, a bowl.
3. A table with a variety of food items, a cake, and a bowl of salad.
4. A person is sitting at a table with a **guitar**.
5. There are also several cups and a knife.

Prediction: guitar

Question: What is the green item with seeds on the plate? GT Answer: pickle

ReCoCap captions:
1. A plate of food with a sandwich and salad on it.
2. A white plate with a sandwich cut in half, accompanied by a side of salad.
3. The sandwich appears to be a tasty-looking grilled cheese sandwich.
4. The sandwich is cut in half, and there are **pickles** on the side.
5. A dining table with a white plate containing a sandwich and a side of salad.

Prediction: pickle

Fig. B.13. Additional qualitative results on OK-VQA.

B.2. Factual content

Fig. B.14 shows additional qualitative results on VQAv2 [52]. The generated descriptions not only capture fine-grained details but also accurately reflect factual content of the images.

Question: What color are the gym shoes? **GT Answer: white**

Initial description :

- A shoe rack filled with various pairs of shoes and sandals.

ReCoD descriptions :

- A shoe rack with a dog sleeping on top of it.
- The dog is curled up and appears to be comfortable in its resting spot.
- A shelf with a dog sleeping on top of it, surrounded by various pairs of shoes.
- A brown puppy sleeping on a shelf next to a **white** sneakers.

Prediction: white

Question: What color are the pillows? **GT Answer: white**

Initial description :

- A bed with a blue comforter and a quilt on top of it.

ReCoD descriptions :

- The bed in the image is a double bed with a blue comforter and **white** pillows.
- The quilt is a patchwork design, adding a cozy and personal touch to the bed.
- The bed is covered with a colorful quilt, which adds a vibrant touch to the room.
- The sheets are pulled down, revealing the blue blanket and the **white** pillows.

Prediction: white

Fig. B.14. Additional qualitative results on VQAv2, demonstrating that the generated descriptions include factual content of the images.

B.3. Object diversity

Fig. B.15 presents qualitative results comparing the object lists extracted from the descriptions generated by each method. Consistent with Fig. 11, the lists produced by RECOD are richer and more diverse. Notably, they include the words “face” and “eyes”, which are taken from the face-shaped alarm clock. They also captures the “city lights” visible outside the window.

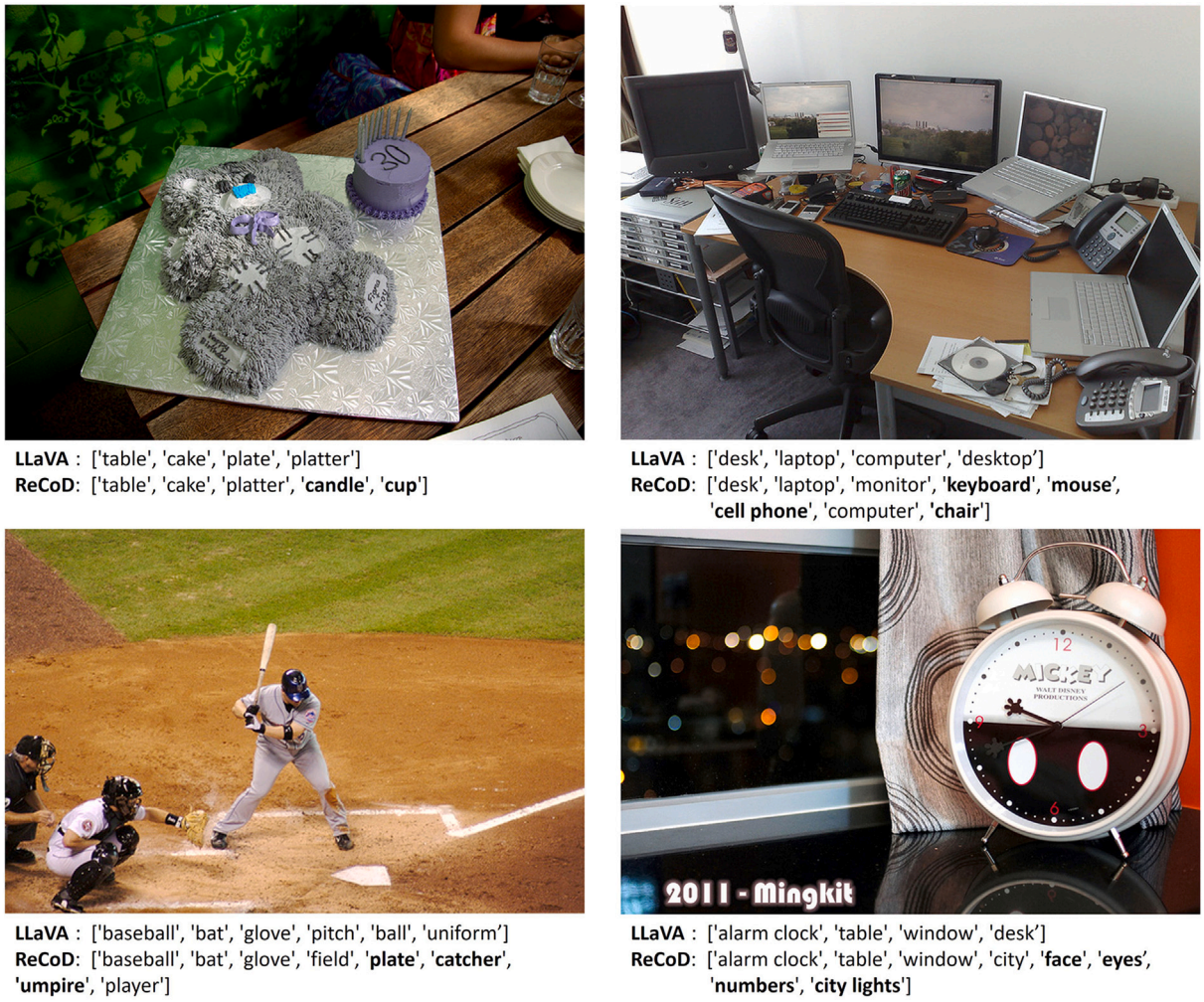


Fig. B.15. Additional qualitative results of object lists extracted from descriptions. RECOD contains more diverse and detailed objects than LLaVA.

Data availability

The data that has been used is publicly available.

References

[1] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.
[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, et al., LLaMA: open and efficient foundation language models, arXiv preprint arXiv:2302.13971, 2023.
[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog (2019).
[5] C. Wei, C. Liu, S. Qiao, Z. Zhang, A. Yuille, J. Yu, De-diffusion makes text a strong cross-modal interface, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13492–13503.
[6] R. Mokady, A. Hertz, ClipCap: CLIP prefix for image captioning, arXiv preprint arXiv:2111.09734, 2021.
[7] Z. Luo, Z. Hu, Y. Xi, R. Zhang, J. Ma, I-tuning: tuning frozen language models with image for lightweight image captioning, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.
[8] H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416, 2022.
[9] AI@Meta, LLaMA 3 model card, 2024, https://github.com/meta-llama/llama3/blob/main.
[10] OpenAI, GPT-4 technical report, arXiv preprint arXiv:2303.08774, 2024.
[11] M. Tsimpoukelli, J. Menick, S. Cabi, S.M.A. Eslami, O. Vinyals, F. Hill, Multimodal few-shot learning with frozen language models, in: Advances in Neural Information Processing Systems, 2021, pp. 200–212.
[12] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, et al., Flamingo: a visual language model for few-shot learning, in: Advances in Neural Information Processing Systems, 2022, pp. 23716–23736.
[13] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, et al., InstructBLIP: towards general-purpose vision-language models with instruction tuning, in: Advances in Neural Information Processing Systems, 2023, pp. 49250–49267.
[14] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, et al., Language is not all you need: aligning perception with language models, in: Advances in Neural Information Processing Systems, 2023, pp. 72096–72109.
[15] H. Liu, C. Li, Y. Li, Y.J. Lee, Improved baselines with visual instruction tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26296–26306.

[16] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, J.-R. Wen, Evaluating object hallucination in large vision-language models, in: Conference on Empirical Methods in Natural Language Processing, 2023, pp. 292–305.

[17] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, et al., Scaling up visual and vision-language representation learning with noisy text supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 4904–4916.

[18] N. Mu, A. Kirillov, D. Wagner, S. Xie, SLIP: self-supervision meets language-image pre-training, in: European Conference on Computer Vision (ECCV), 2022, pp. 529–544.

[19] L. Yao, R. Pi, J. Han, X. Liang, H. Xu, W. Zhang, et al., DetCLIPv3: towards versatile generative open-vocabulary object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27391–27401.

[20] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, Y. Shan, YOLO-world: real-time open-vocabulary object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16901–16911.

[21] S. Fu, Q. Yang, Q. Mo, J. Yan, X. Wei, J. Meng, et al., LLMDet: learning strong open-vocabulary object detectors under the supervision of large language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 14987–14997.

[22] S. Ma, Y. Wang, Y. Wei, E. Zhang, J. Fan, X. Sun, et al., SKDF: a simple knowledge distillation framework for distilling open-vocabulary knowledge to open-world object detector, IEEE Trans. Pattern Anal. Mach. Intell. (2025) 1–16.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Computer Vision – ECCV 2020, 2020, pp. 213–229.

[24] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: deformable transformers for end-to-end object detection, in: International Conference on Learning Representations, 2021.

[25] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464–7475.

[26] F. Zuo, J. Liu, Z. Chen, H. Zhang, M. Fu, L. Wang, Multilevel fine-grained features-based general framework for object detection, IEEE Trans. Cybern. 54 (11) (2024) 6921–6933.

[27] B. Li, K.Q. Weinberger, S. Belongie, V. Koltun, R. Ranftl, Language-driven semantic segmentation, in: International Conference on Learning Representations, 2022.

[28] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, et al., DenseCLIP: language-guided dense prediction with context-aware prompting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18082–18091.

[29] C. Zhou, C.C. Loy, B. Dai, Extract free dense labels from CLIP, in: European Conference on Computer Vision (ECCV), 2022, pp. 696–712.

[30] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, et al., Scaling up vision-language pre-training for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17980–17989.

[31] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, et al., GIT: a generative image-to-text transformer for vision and language, Trans. Mach. Learn. Res. (2022).

[32] R. Dessì, M. Bevilacqua, E. Gualdoni, N.C. Rakotonirina, F. Franzon, M. Baroni, Cross-domain image captioning with discriminative finetuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6935–6944.

[33] W. Li, L. Zhu, L. Wen, Y. Yang, DeCap: decoding CLIP latents for zero-shot captioning via text-only training, in: The Eleventh International Conference on Learning Representations, 2023.

[34] R. Ramos, B. Martins, D. Elliott, Y. Kementchedjhieva, SmallCap: lightweight image captioning prompted with retrieval augmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2840–2849.

[35] Z. Zeng, Y. Xie, H. Zhang, C. Chen, B. Chen, Z. Wang, MeaCap: memory-augmented zero-shot image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14100–14110.

[36] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models, in: Proceedings of the 40th the International Conference on Machine Learning, PMLR, 2023, pp. 19730–19742.

[37] H. Liu, C. Li, Q. Wu, Y.J. Lee, Visual instruction tuning, in: Advances in Neural Information Processing Systems, 2023, pp. 34892–34916.

[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[39] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM (2014) 78–85.

[40] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: an open multilingual graph of general knowledge, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017, pp. 4444–4451.

[41] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, et al., An empirical study of GPT-3 for few-shot knowledge-based VQA, Proc. AAAI Conf. Artif. Intell. (2022) 3081–3089.

[42] Y. Du, J. Li, T. Tang, W.X. Zhao, J.-R. Wen, Zero-shot visual question answering with language model feedback, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 9268–9281.

[43] J. Guo, J. Li, D. Li, A.M.H. Tiong, B. Li, D. Tao, et al., From images to textual prompts: zero-shot visual question answering with frozen large language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10867–10877.

[44] Y. Hu, H. Hua, Z. Yang, W. Shi, N.A. Smith, J. Luo, PromptCap: prompt-guided image captioning for VQA with GPT-3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2963–2975.

[45] Z. Shao, Z. Yu, M. Wang, J. Yu, Prompting large language models with answer heuristics for knowledge-based visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14974–14983.

[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft COCO: common objects in context, in: European Conference on Computer Vision (ECCV), 2014, pp. 740–755.

[47] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565.

[48] A. Burns, K. Srinivasan, J. Ainslie, G. Brown, B.A. Plummer, K. Saenko, et al., WikiWeb2M: a page-level multimodal wikipedia dataset, arXiv preprint arXiv:2305.05432, 2023.

[49] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, 2020, pp. 1877–1901.

[50] K. Marino, M. Rastegari, A. Farhadi, R. Mottaghi, Ok-VQA: a visual question answering benchmark requiring external knowledge, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3195–3204.

[51] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, R. Mottaghi, A-OKVQA: a benchmark for visual question answering using world knowledge, in: European Conference on Computer Vision (ECCV), 2022, pp. 146–162.

[52] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the V in VQA matter: elevating the role of image understanding in visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6904–6913.

[53] Anthropic, The claude 3 model family: opus, sonnet, haiku, Anthropic Blog, 2024.

[54] F. Gardères, M. Ziaeefard, B. Abeloos, F. Lecue, ConceptBert: concept-aware representation for visual question answering, in: Conference on Empirical Methods in Natural Language Processing, 2020, pp. 489–498.

[55] K. Marino, X. Chen, D. Parikh, A. Gupta, M. Rohrbach, KRISP: integrating implicit and symbolic knowledge for open-domain knowledge-based VQA, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14111–14121.

[56] J. Wu, J. Lu, A. Sabharwal, R. Mottaghi, Multi-modal answer validation for knowledge-based VQA, Proc. AAAI Conf. Artif. Intell. (2022) 2712–2721.

[57] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, J. Gao, KAT: a knowledge augmented transformer for vision-and-language, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 956–968.

[58] L. Lin, Y. Xie, D. Chen, Y. Xu, C. Zhu, L. Yuan, REVIVE: regional visual representation matters in knowledge-based visual question answering, in: Advances in Neural Information Processing Systems, 2022, pp. 10560–10571.

[59] A.M.H. Tiong, J. Li, B. Li, S. Savarese, S.C.H. Hoi, Plug-and-play VQA: zero-shot VQA by conjoining large pretrained models with zero training, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 951–967.

[60] J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Advances in Neural Information Processing Systems, 2019, pp. 13–23.

[61] H. Tan, M. Bansal, 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), 2019, pp. 5100–5111.

[62] A. Kamath, C. Clark, T. Gupta, E. Kolve, D. Hoiem, A. Kembhavi, Webly supervised concept expansion for general purpose vision models, in: European Conference on Computer Vision (ECCV), 2022, pp. 662–681.

[63] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, et al., Qwen2-VL: enhancing vision-language model's perception of the world at any resolution, arXiv preprint arXiv:2409.12191, 2024.

[64] Y. Li, C. Wang, J. Jia, LLaMA-VID: an image is worth 2 tokens in large language models, in: European Conference on Computer Vision (ECCV), 2024, pp. 323–340.

[65] S. Yang, Y. Chen, Z. Tian, C. Wang, J. Li, B. Yu, et al., VisionZip: longer is better but not necessary in vision language models, in: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19792–19802.

[66] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[67] S. Banerjee, A. Lavie, METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.

[68] C.-Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[69] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566–4575.
[70] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: semantic propositional image caption evaluation, in: European Conference on Computer Vision (ECCV), 2016, pp. 382–398.
[71] A. Hurst, A. Lerer, A.P. Goucher, A. Perelman, A. Ramesh, A. Clark, et al., GPT-4o system card, arXiv preprint arXiv:2410.21276, 2024.
[72] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, J. Zou, Mind the gap: understanding the modality gap in multi-modal contrastive representation learning, in: Advances in Neural Information Processing Systems, 2022, pp. 17612–17625.

Author biography

Geunyoung Jung received the B.S. degree in Physics from University of Seoul, Seoul, South Korea in 2022. He is currently pursuing the PhD degree in Artificial Intelligence at the University of Seoul. His research interests include 3D vision, multimodal learning, and their practical applications in real-world scenarios.

Jun Park received the B.S. degree in Artificial Intelligence from University of Seoul, Seoul, South Korea in 2025. He is currently pursuing the master's degree in robotics program at the KAIST. His research interests include the intersection of deep learning and computer vision, with particular emphasis on their applications in robotics.

Hankyeol Lee received the M.S. degree in Artificial Intelligence from University of Seoul, Seoul, South Korea in 2025. He is currently working as a research engineer at KT. His research interest includes multimodal learning, specifically in vision-language models.

Kyungwoo Song worked as an assistant professor in the Department of Artificial Intelligence at the University of Seoul, South Korea, from 2021 to 2023. Since 2023, he has been with the Department of Statistics and Data Science, Yonsei University, South Korea.

Jiyoung Jung received the BS, MS, and PhD degrees in Electrical Engineering from KAIST, South Korea in 2008, 2010, and 2016, respectively. She worked as a research engineer in the mobility team at Naver Labs in 2016, then she worked as an assistant professor in the Department of Software Convergence at Kyung Hee University, South Korea from 2017 to 2021. Since 2021, she has been a faculty member in the Department of Artificial Intelligence at the University of Seoul, South Korea. Her research interests include 3D modeling, multi-sensor systems, and vision-language models.