# RoRA-VLM: Robust Retrieval Augmentation for Vision Language Models

Anonymous ACL submission

## Abstract

Recent vision-language models (VLMs), despite their broad capabilities, continue to underperform on knowledge-intensive tasks. Retrieval augmentation offers a promising solution by incorporating external multimodal knowledge. However, the retrieved content often contains a mix of relevant and irrelevant information, and existing methods primarily focus on improving retrieval quality to mitigate this issue. In this work, we propose RoRA-VLM, a robust retrieval augmentation framework designed to address the complementary challenge of utilizing noisy retrieved knowledge effectively. The core insight behind RoRA-VLM is that the multimodal nature of VLMs enables a novel solution: visual information can act as a signal for assessing the relevance of retrieved results. To this end, RoRA-VLM introduces a learned cross-modal verification mechanism that enables VLMs to compare visual similarities between the query and retrieved images, and attend selectively to visually relevant retrievals while filtering out irrelevant content. Extensive experiments on OVEN, InfoSeek, and Enc-VQA benchmarks demonstrate that RoRA-VLM achieves significant performance improvements of up to 14.76% in accuracy compared to baseline models with minimal training data, consistently outperforming state-of-the-art retrieval-augmented VLMs while exhibiting strong generalization to unseen domains.

## 1 Introduction

Vision-language models (VLMs) (Li et al., 2023; Alayrac et al., 2022; Liu et al., 2023b; Dai et al., 2023) have achieved remarkable progress across various visual perception and generation tasks (Antol et al., 2015; Marino et al., 2019; Dai et al., 2024). However, despite these advancements, recent studies (Chen et al., 2023d; Hu et al., 2023; Mensink et al., 2023) reveal that VLMs still face significant challenges in knowledge-intensive tasks, such as visual entity grounding (Hu et al., 2023) and information-seeking visual question answering (Chen et al., 2023d), where VLMs need to effectively link visual objects and scenes to their corresponding entities and relevant background knowledge. For instance, as illustrated in Figure 1, given the question "Who designed the tallest building in the picture?" alongside an image of several buildings, VLMs need to accurately identify the building based on its visual attributes and retrieve the associated background knowledge. The vast and dynamic nature of visual knowledge in the open world, however, makes it impractical for VLMs to store all possible associations between visual appearances and their corresponding entities and background knowledge in their parameters.

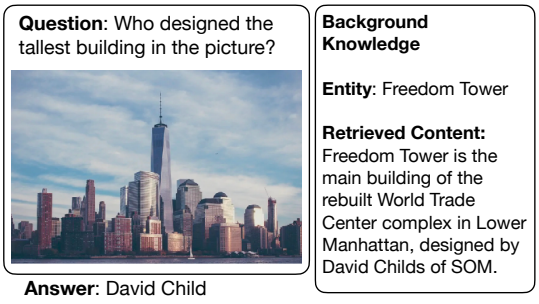


Figure 1: An example question for information-seeking visual question answering.

Retrieval-augmented generation (RAG) offers a promising solution by integrating knowledge retrieved from external sources with VLMs, demonstrating success in improving text-based knowledge-intensive tasks for LLMs (Guu et al., 2020; Lewis et al., 2020; Yoran et al., 2023). However, incorporating retrieval augmentation introduces a fundamental challenge: assessing the relevance of retrieved information. While prior studies on robust RAG (Yoran et al., 2023) have explored techniques to make language models resilient to noisy retrievals in text-only tasks, these approaches

overlook a unique advantage in the multimodal setting: the ability to leverage information from different modalities as evidence for retrieved knowledge relevance evaluation. In multimodal contexts, when a retrieved passage accompanies a corresponding image, the model can compare visual similarities between the query image and retrieved images to determine which retrieved knowledge is most likely relevant to the query, offering an explicit verification mechanism for filtering and prioritizing the retrieved information.

In this paper, we introduce RoRA-VLM, a robust retrieval-augmented framework that leverages visual information as evidence for evaluating the quality of retrieved knowledge. Unlike previous approaches that focus primarily on improving retrieval quality, our work addresses the complementary challenge of how to effectively utilize retrieved knowledge. Our framework implements this vision-guided relevance assessment with two primary components, as shown in Figure 2. First, we develop a **Multimodal-Reciprocal Retrieval** that integrates cross-modal retrieval with visual token refinement to acquire external high-quality multimodal knowledge. Second, we formulate a **Cross-Modal Verification Mechanism** that enables VLMs to perform effective verification between modalities. This mechanism is realized through a two-phase training: a knowledge-intensive pre-training followed by an adversarial noise injection fine-tuning, enabling VLMs to effectively distinguish relevant knowledge from noise.

We conduct extensive experiments to evaluate the effectiveness and robustness of our proposed framework on three widely adopted knowledge-seeking benchmarks: OVEN (Hu et al., 2023), InfoSeek (Chen et al., 2023d), and Enc-VQA (Mensink et al., 2023). Our results demonstrate that, with only a minimal number of training instances (e.g., 10,000), the framework achieves significant improvements over baseline models, yielding up to 14.36% accuracy improvement, and consistently outperforms the current SOTA retrieval-augmented generation methods. Additionally, our analysis explicitly reveals that RoRA-VLM effectively learns to discriminate between relevant and irrelevant retrieved information by comparing visual similarities, demonstrates significant robustness to retrieval noise, and shows strong zero-shot transfer capabilities to knowledge-intensive tasks from unseen domains.

## 2 Related Work

**Vision-Language Models** Recent advancements in vision-language models (VLMs), such as BLIP-2 (Li et al., 2023), Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023b), and InstructBLIP (Dai et al., 2023), have demonstrated remarkable performance on various visual perception tasks, such as image captioning (Lin et al., 2014; Schuhmann et al., 2022; Chen et al., 2023a), visual question answering (Antol et al., 2015; Marino et al., 2019; Schwenk et al., 2022), object detection (Lin et al., 2014; Everingham et al.), visual grounding (Hu et al., 2023; Kazemzadeh et al., 2014), and visual relationship detection (Lu et al., 2016), etc. These models typically employ an architecture consisting of a pre-trained visual encoder (Radford et al., 2021; Dosovitskiy et al., 2021; Chen et al., 2024), a pre-trained large language model (Touvron et al., 2023; Almazrouei et al., 2023), and a projection function that maps visual features to the text embedding space (Liu et al., 2023b). However, this method often falls short in aligning visual features with the extensive knowledge embedded in language models. Alternative architectures, such as the Q-former used in BLIP-2 (Li et al., 2023) and the perceiver resampler in Flamingo (Alayrac et al., 2022), have been proposed to enhance the perception of visual content.

**Knowledge-Intensive Tasks and Retrieval-Augmented Generation** Augmenting models with external knowledge sources has proven effective in enhancing performance on knowledge-intensive tasks. In the text-only domain, models like REALM (Guu et al., 2020), RAG (Lewis et al., 2020), and RobustRAG (Yoran et al., 2023) have demonstrated the benefits of retrieval-based augmentation by providing additional context for generating accurate responses. Applying this approach to the vision-language domain presents unique challenges due to modality discrepancies and differing model architectures (Wei et al., 2023). Recent studies (Gui et al., 2021; Lin et al., 2023, 2024) have explored multimodal retrieval to enhance LLMs by retrieving textual knowledge from visual queries, but primarily focus on improving retrieval quality rather than effectively utilizing retrieved information. Given that state-of-the-art retrievers achieve only modest performance (e.g., lower than 0.2 for recall@1 on InfoSeek (Chen et al., 2023d)), managing and denoising retrieval noise becomes crucial for VLMs. Knowledge-intensive bench-

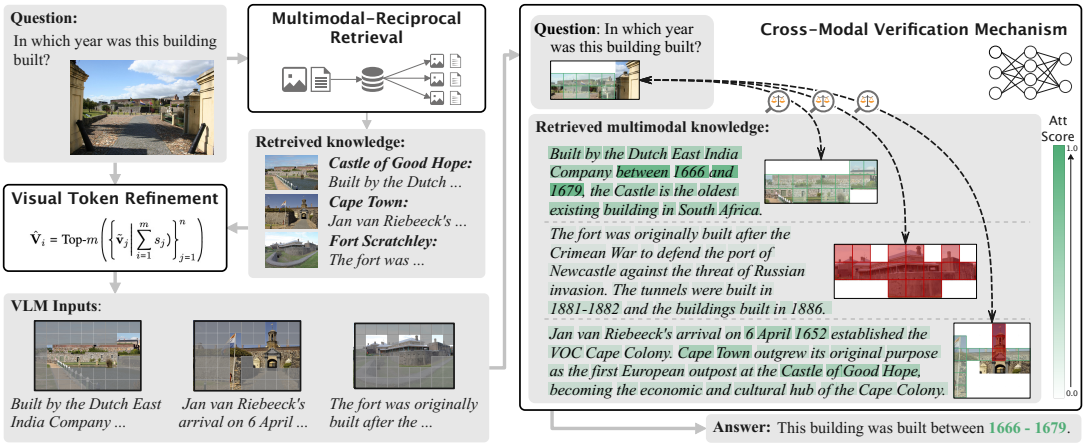


Figure 2: Overview architecture of RoRA-VLM. Our approach teaches VLMs to use visual information as evidence for evaluating the quality of retrieved knowledge through cross-modal verification. By comparing visual similarities between the query image and retrieved images, the model learns to selectively attend to information from relevant retrievals while ignoring irrelevant ones.

marks such as OVEN (Hu et al., 2023) and InfoSeek (Chen et al., 2023d) have been developed to evaluate VLMs on tasks like visual entity grounding and information-seeking visual question answering, which require models to recognize visual entities and connect them with background knowledge. Studies have shown that extensive fine-tuning on knowledge-intensive task instances does not substantially improve VLMs' performance (Chen et al., 2023d; Hu et al., 2023; Mensink et al., 2023), indicating that current architectures struggle with dynamic visual-semantic associations.

## 3 RoRA-VLM

In this work, we focus on improving VLMs on knowledge-intensive VQA tasks via retrieval-augmented generation. Given a text query $t$ together with an image $v$, a VLM is expected to generate a response $y$ by leveraging the multimodal knowledge snippets $R$ retrieved from an external database as context. The objective of the retrieval-augmented generation can be formulated as:

$$y = \arg\max_{y} P(y|t, v, R) \quad (1)$$

Our proposed framework, RoRA-VLM, enables VLMs to robustly leverage retrieved knowledge through vision-guided relevance assessment. The core insight is training VLMs to use visual information as evidence for evaluating the quality of retrieved knowledge, allowing them to determine which retrievals are most relevant to the query and filter out noise. This capability is achieved through two components: **(1) Multimodal-Reciprocal Retrieval**, which acquires high-quality multimodal knowledge with visual cues guided textual retrieval and entity-based visual token refinement, and **(2) Cross-Modal Verification Mechanism**, which enables VLMs to perform verification across different modalities to distinguish relevant knowledge from irrelevant information. The overall architecture of RoRA-VLM is shown in Figure 2.

### 3.1 Multimodal-Reciprocal Retrieval

Traditional retrieval methods face unique challenges with vision-language tasks due to inherent modality discrepancies. In multimodal settings, textual queries often contain generic terms or anaphoric references (e.g., "the tallest building") that lack specificity without visual context, while visual information alone may not sufficiently clarify the query's intent. We overcome these challenges by developing a multimodal-reciprocal retrieval method that leverages visual cues to guide textual retrieval, followed by a visual feature refinement based on retrieved textual entity information.

**Image-Anchored Entity Retrieval** Given the textual query $t$ and query image $v$, we first use the query image $v$ as an anchor to retrieve top-$k$ most similar images from an image database built upon WIT (Srinivasan et al., 2021). This image-level retrieval identifies potential entities that visually match the query image, providing crucial visual evidence for subsequent cross-modal verification. For efficient retrieval, we encode all images in the WIT dataset with a CLIP (Radford

et al., 2021) image encoder and construct a dense vector-search database[1]. Given a query image, the image retriever $\phi^{\text{v}}$ computes cosine similarities between the query image feature and retrieval database keys to fetch the top-$k$ most similar images $\tilde{V} = \{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k\}$ along with their associated entity names and background information $\tilde{E} = \{\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_k\}$. More details of the CLIP-based image retriever are provided in Appendix B.

**Query-Expanded Text Retrieval** With the entity names and descriptions obtained from the image-anchored entity retrieval, we further use them to expand the original text query. This expansion disambiguates anaphoric references in the original query by providing specific entity information. Given the original text query $t$ and a retrieved entity name and description $e_i$, the text retriever $\phi^{\text{t}}$ [2] searches for top-$l$ textual knowledge snippets that are most relevant to the expanded query $\tilde{C}_i = \{\tilde{c_{i,1}}, \tilde{c_{i,2}}, \ldots, \tilde{c_{i,l}}\} = \phi^{\text{t}}(t, \tilde{e}_i)$. We then concatenate each retrieved image $\tilde{v}_i$ with its corresponding textual knowledge snippets $C_i$ to form multimodal knowledge snippets $R = \{r_1, r_2, \ldots, r_k\} = \{[\tilde{v}_1, \tilde{e}_1, \tilde{C}_1], [\tilde{v}_2, \tilde{e}_2, \tilde{C}_2], \ldots, [\tilde{v}_k, \tilde{e}_k, \tilde{C}_k]\}$.

**Visual Token Refinement** Retrieved images often contain noise, such as objects or visual scenes that are irrelevant to the concerned entities, which can distract the model during subsequent cross-modal verification. To enhance the quality of visual representations by focusing only on the most query-relevant features, we implement a refinement strategy that filters out irrelevant visual information. Specifically, for the query image, we first select the top-$m$ visual tokens most similar to the text query embedding:

$$\hat{\mathbf{V}} = \text{Top-}m\left(\left\{\mathbf{v}_i \middle| s_i\right\}_{i=1}^{n}\right), \tag{2}$$

where $s_i = \mathbf{v}_i \cdot \mathbf{t}$ represents the similarity between visual token $\mathbf{v}_i$ and text query embedding $\mathbf{t}$, and $n$ represents the total number of visual tokens extracted from each image by the vision encoder. Subsequently, for each retrieved image, we select the top-$m$ visual tokens most similar to the refined query image tokens:

$$\hat{\mathbf{V}}_i = \text{Top-}m\left(\left\{\tilde{\mathbf{v}}_j \middle| \sum_{i=1}^{m} s_j\right\}_{j=1}^{n}\right). \tag{3}$$

where $s_j = \sum_{i=1}^{m}(\mathbf{v}_i \cdot \tilde{\mathbf{v}}_j)$ with $\mathbf{v}_i \in \hat{\mathbf{V}}$. This refinement process ensures that the model focuses on the most query-relevant visual features when performing cross-modal verification, enabling more accurate assessment of retrieval relevance. More details of the visual token refinement are provided in Appendix C.

### 3.2 Cross-Modal Verification Mechanism

The second key component of our framework is a Cross-Modal Verification Mechanism that evaluates the relevance of retrieved knowledge. We enable VLMs to implicitly leverage this mechanism through a two-phase training: (1) a knowledge-intensive pre-training to align visual features with LLMs' internal knowledge, followed by (2) an adversarial noise injection fine-tuning that enables discrimination between relevant and irrelevant information through contrastive learning signals.

**Knowledge-Intensive Pre-training** The ability to compare visual entities across images requires a strong foundation in visual-knowledge alignment. To establish this foundation, we conduct pre-training on WikiWeb2M (Burns et al., 2023), a knowledge-intensive multimodal dataset. We curate 1 million entity-rich image-text instances, each consisting of an image depicting an entity, its caption, and associated textual knowledge. This pre-training phase aligns visual appearances with entity knowledge stored in the LLM, enhancing the model's ability to recognize the same entity across different visual representations and contexts.

**Adversarial Noise Injection Fine-tuning** An effective training for cross-modal verification mechanism learning requires exposing the model to both relevant and irrelevant retrievals. Since ground-truth relevance labels are not accessible for our training data, we employ an adversarial noise injection fine-tuning strategy to create a controlled learning environment with clear contrastive signals.

For each training instance $(t, v)$, we retrieve the top-($k$-1) multimodal knowledge snippets $R = r_1, r_2, \ldots, r_{k-1}$. We then deliberately introduce an irrelevant knowledge snippet $r' = [\tilde{v}', \tilde{e}', \tilde{C}']$ randomly sampled from the retrieval database. This creates a contrastive learning signal that facilitates

[1] We construct the vector-search database based on a hierarchical navigable small-world (HNSW) graph (Malkov and Yashunin, 2018).

[2] The textual retrieval component is implemented using Google Search functionality accessed through the Serper API service: `https://serper.dev/`.

the model's ability to distinguish between relevant and irrelevant information. The input is formed by concatenating these snippets with the original query: $[r_1 : r_2 : \ldots : r' : t : v]$. We fine-tune VLMs on such noise-injected training instances by minimizing the cross-entropy loss of predicting the target answer despite the presence of irrelevant information.

## 4 Experiment Setup

**Evaluation Benchmarks** To evaluate the effectiveness and robustness of RoRA-VLM, we conduct experiments on three benchmark datasets, including OVEN (Hu et al., 2023) for visual entity grounding, and InfoSeek (Chen et al., 2023d) and Encyclopedic-VQA (Mensink et al., 2023) for information-seeking visual question answering. As the test sets of OVEN and InfoSeek are not available at the time of submission, we report our results on their validation sets.

**Evaluation Metrics** We adopt evaluation metrics in line with previous studies (Hu et al., 2023; Chen et al., 2023d; Mensink et al., 2023). For visual entity recognition tasks, we use the standard *accuracy* metric to assess the model's capability to correctly identify entities in images. For knowledge-seeking visual question answering (VQA) tasks, we apply different metrics tailored to specific question types. For questions expecting string-based responses such as entity names, we report accuracy using the *VQA accuracy* metric (Antol et al., 2015). For questions requiring numeric answers, we use *relaxed accuracy* (Methani et al., 2020), which accounts for deviations from exact numerical values.

**Baselines** We compare our framework with several state-of-the-art vision-language models, including LLaVA-v1.5/v1.6 (Liu et al., 2023a, 2024), PaLI-17B/X (Chen et al., 2023c,b), BLIP-2/InstructBLIP (Li et al., 2023; Dai et al., 2023), CLIP2CLIP (Hu et al., 2023), and Qwen-2.5-VL 3B. We also include retrieval-augmented approaches such as PreFLMR (Lin et al., 2024), RA-CM3 (Yasunaga et al., 2023), and Wiki-LLaVA (Caffagni et al., 2024), which leverage external knowledge to enhance generation. To ensure a fair comparison, all baseline models are fine-tuned on the OVEN (Hu et al., 2023), InfoSeek (Chen et al., 2023d), and Enc-VQA (Mensink et al., 2023) datasets respectively, and then evaluated on the corresponding tasks.

| Model | OVEN | | InfoSeek | | Enc-VQA |
|---|---|---|---|---|---|
| | Entity | Query | Entity | Query | |
| *Non Retrieval-Augmented Model* | | | | | |
| CLIP2CLIP | 10.10 | 2.10 | - | - | - |
| PaLI | 12.40 | 22.40 | 16.00 | 20.70 | - |
| PaLI-X | - | - | 20.80 | 23.50 | - |
| BLIP-2 | - | - | 13.30 | 14.50 | - |
| InstructBLIP | - | - | 13.20 | 14.30 | - |
| LLaVA-v1.6 | 3.72 | 24.55 | 14.16 | 15.98 | 13.54 |
| LLaVA-v1.5 | 3.63 | 20.04 | 10.34 | 12.98 | 12.21 |
| Qwen-2.5-VL | 16.30 | 44.26 | 19.66 | 22.50 | 13.53 |
| *Retrieval-Augmented Model* | | | | | |
| RA-CM3 | - | - | 17.09 | 21.64 | - |
| PreFLMR | - | - | 19.37 | 22.21 | - |
| Wiki-LLaVA* | 14.43 | 20.4 | 21.44 | 23.68 | 18.61 |
| RoRA-VLM | | | | | |
| - LLaVA-1.5 | 15.08 | 24.06 | **25.10** | **27.34** | 20.29 |
| - Qwen-2.5-VL | **27.51** | **51.20** | 23.96 | 26.27 | **20.42** |

Table 1: Evaluation results in accuracy (%). The best performance is highlighted in **bold**. The Entity groups expect an entity name as the target answer, while Query groups target a general object name or concept as the answer. * denotes our implementation of Wiki-LLaVA as its original source code is not publicly available.

**Model Tuning** We demonstrate the general applicability of our approach across different architectures by implementing it on multiple backbone models, including LLaVA-v1.5 7B (Liu et al., 2023a) and Qwen-2.5-VL 3B (Bai et al., 2023). For comparative analysis, we apply the complete framework to LLaVA-v1.5, while implementing RoRA-VLM on Qwen-2.5-VL 3B without the knowledge-intensive pre-training phase to isolate direct performance improvements. Our experimental validation uses 10,000 randomly sampled instances from each benchmark dataset (OVEN, InfoSeek, Encyclopedic-VQA) for efficient fine-tuning.

## 5 Result & Discussion

### 5.1 Main Results

Table 1 presents the main results for visual entity grounding on the OVEN dataset and information-seeking visual question answering on the InfoSeek and Encyclopedic-VQA datasets. Despite being fine-tuned on only 10,000 instances per dataset, both RoRA variants show remarkable gains. When applied to LLaVA-v1.5, our approach yields substantial improvements over the base model (e.g., from 10.34% to 25.10% on InfoSeek-Entity). Similarly, when applied to Qwen-2.5-VL, our approach boosts performance across all benchmarks (e.g., from 16.30% to 27.51% on OVEN-Entity). In addition, our approach consistently outperforms previous retrieval-augmented methods, such as RA-CM3, PreFLMR, and Wiki-LLaVA, demonstrating

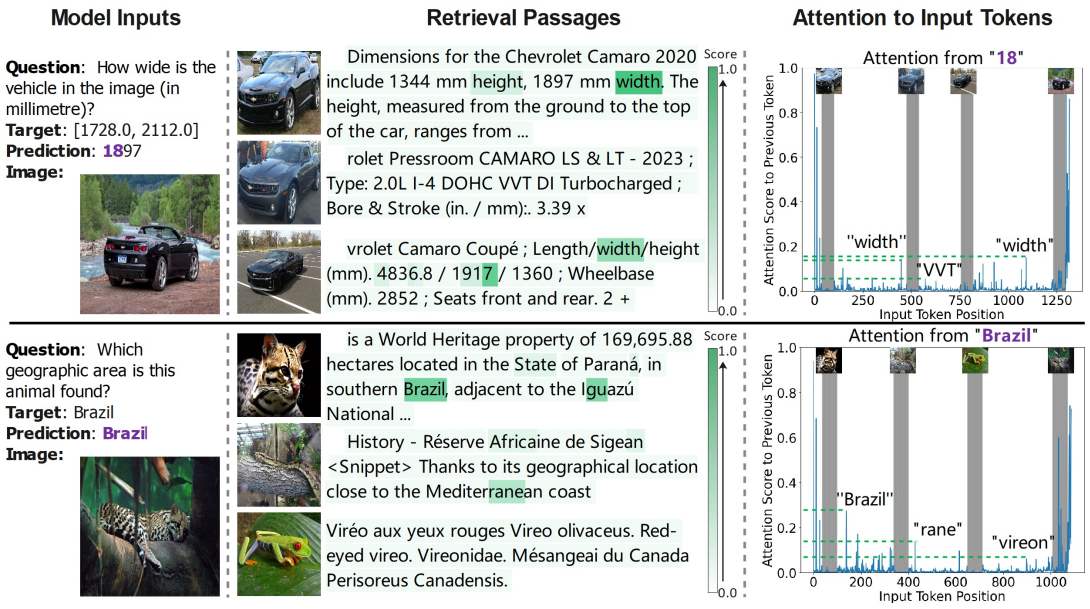


Figure 3: Visualization of attention scores assigned to VLM input tokens during next-token generation. Tokens are highlighted in green, with darker shades indicating higher attention scores.

the advantage of cross-modal verification mechanism in handling retrieval noise.

## 5.2 Effect of Cross-Modal Verification Mechanism

We conducted two ablation studies to validate the effectiveness of our cross-modal verification approach. First, we implemented a "textual-only RAG" configuration by removing retrieved images while keeping all other aspects identical, thus eliminating the model's ability to perform visual comparison between query and retrieved content. As shown in Table 2, this modification resulted in a substantial performance decrease of 7.81-8.06 percentage points across benchmarks, confirming that visual modality provides crucial evidence for determining relevance that cannot be obtained from text alone. Second, we examined a variant where we maintained the multimodal retrievals but removed the adversarial noise injection during training. This configuration exhibited an even more significant performance decline of 8.51-9.92 percentage points compared to our full approach, as presented in Tabel 2, demonstrating that training using only positive samples of correctly retrieved multimodal information is insufficient—the model must be explicitly trained to perform cross-modal verification through contrastive learning with adversarial examples. In addition, we investigate the effect of knowledge-intensive pretraining, demonstrating the importance of alignment between the visual appearances and entity knowledge for knowledge-intensive tasks. More details of the ablation studies of the knowledge-intensive pre-training are provided in Appendix E.1.

| Model | Entity | Query |
|---|---|---|
| LLaVA-1.5 | 10.34 | 12.98 |
| RoRA LLaVA-1.5 | **25.10** | **27.34** |
| *Visual Token Refinement Ablations* | | |
| - w/o visual token refinement (pooling) | 23.94 | 24.85 |
| - w/o visual token refinement (all tokens) | 24.62 | 26.14 |
| *Cross-Modal Verification Ablations* | | |
| - w/o retrieved images (textual-only RAG) | 17.29 | 19.28 |
| - w/o noise-injection | 16.59 | 17.42 |

Table 2: Ablation studies of each component in RoRA-VLM. Performance is reported in accuracy (%) on InfoSeek.

To further demonstrate that the model has learned this cross-modal verification capability, we visualize the attention scores assigned to tokens of the retrieved knowledge during the answer generation in Figure 3. The visualization reveals that RoRA-VLM effectively learns to focus on textual knowledge corresponding to images containing entities that match those in the query image. For instance, in the second row of Figure 3, the model predominantly attends to the first two knowledge snippets while disregarding the third, which pertains to a completely different animal. This attention pattern directly evidences that the RoRA-VLM has learned to leverage visual similarity as a signal for relevance assessment, enabling effective noise filtering in the retrieval augmentation

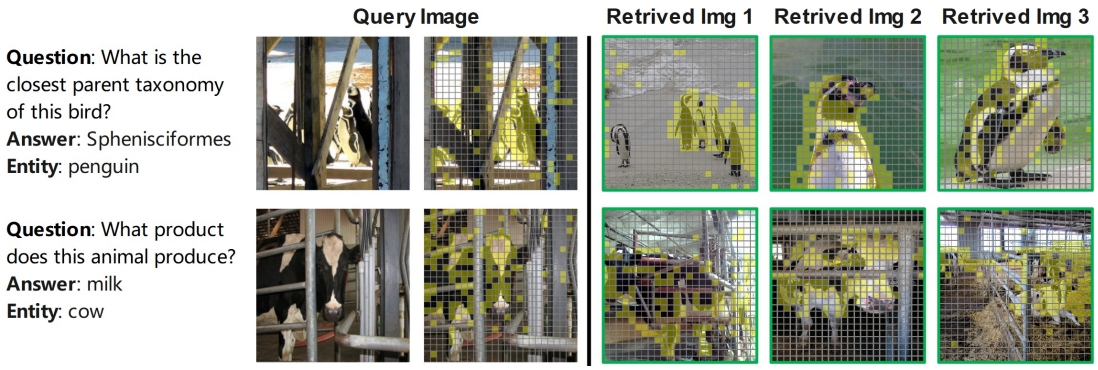


Figure 4: Qualitative results for query-oriented Visual Token Refinement .

generation.

### 5.3 Validation of Multimodal-Reciprocal Retrieval

We conducted comprehensive analyses of our Multimodal-Reciprocal Retrieval to evaluate its effectiveness. Table 3 reports the retrieval precision at each stage of our proposed retrieval process. In the image-anchored entity retrieval, we consider retrieval successful if the target entity shown in the query image matches any of the retrieved $m$ images. Similarly, in the query-expanded text retrieval, we consider retrieval successful if the golden answer is included in any of the retrieved textual knowledge snippets.

| Retrieval Stage | OVEN | | InfoSeek | |
|---|---|---|---|---|
| | Entity | Query | Entity | Query |
| Image-Anchored Entity Retrieval | 35.16 | 34.45 | 38.53 | 37.67 |
| Query-Expanded Text Retrieval | - | - | 27.01 | 26.97 |

Table 3: Retrieval precision (%) for the Image-Anchored Entity Retrieval and Query-Expanded Text Retrieval.

To quantify the benefits of our multi-stage approach, we performed an ablation study comparing it with a single-stage retrieval method, and present results in Table 4. In the single-stage configuration, we directly utilized CLIP embeddings of the query image to retrieve similar entity images and their corresponding background knowledge, bypassing the query-expanded text retrieval phase. In addition, we further compare our multi-stage retrieval with previous work, RA-CM3, which employed a single-stage retrieval that utilizes both text and image CLIP embeddings. The experiment results demonstrate that our multimodal-reciprocal retrieval approach consistently outperformed all single-stage retrieval approaches, confirming that the integration of textual queries with visually-derived entities substantially improves retrieval precision. More qualitative examples of the multimodal-reciprocal retrieval could be found in Appendix D.

| Model | Entity | Query |
|---|---|---|
| LLaVA-v1.5 | 10.34 | 12.98 |
| RA-CM3 (single-stage) | 17.09 | 21.64 |
| RoRA-VLM (single-stage) | 21.9 | 23.87 |
| RoRA-VLM (2-stage) | 25.10 | 27.34 |

Table 4: Ablation studies for 2-stage retrieval. Performance is reported in accuracy (%) on InfoSeek.

Complementing the multi-step retrieval, our Visual Token Refinement provides further performance improvements by focusing the model's attention on the most query-relevant visual features. In Figure 4, we show the qualitative results of the visual token refinement . From the query image, we select m=144 visual tokens that are most related to the text query (i.e., the Question), while each visual token corresponds to an image patch (highlighted in yellow). As we can see, this method effectively identifies and selects patches corresponding to the key visual entity, even with the presence of anaphoric references in the query. Similarly, for each retrieved image, we also select m=144 visual tokens that are most related to the query image. These qualitative results underscore the effectiveness of our visual token refinement in filtering out irrelevant visual information, enabling the retrieval augmentation of VLMs more robust. We validate the benefits gained from the visual token refinement through two controlled experiments and listed the experiment results in Table 2. First, replacing our refinement strategy with average pooling to obtain the same number of tokens (144) resulted in a performance decrease of 1.16-2.49 percentage points across benchmarks. This indicates that selecting query-relevant visual features based on semantic similarity is more effective than uniform

dimensionality reduction. Second, we evaluated a variant that uses all 576 visual tokens for each image without refinement. Despite having access to more visual information, this approach still underperformed our method by 0.48-1.20 percentage points while incurring higher computational costs. These results demonstrate that our refinement strategy successfully identifies the most query-relevant visual features, enhancing cross-modal verification accuracy while maintaining computational efficiency. More details of the pooling process for this ablation study are provided in Appendix C.

### 5.4 Robustness to Retrieval Noise

To directly evaluate the robustness of our approach to retrieval noise, we conduct controlled experiments with varying levels of noise injection during inference. Table 5 presents the results on the InfoSeek dataset. We first establish a baseline using only the top-1 retrieved entity image and its corresponding knowledge snippet for generation augmentation. We then create noisy retrieval settings by adding two randomly sampled irrelevant entity images and their knowledge snippets. This random sampling process is repeated twice, resulting in two distinct sets of irrelevant entity images and knowledge snippets for the same input instance. The results show that the model's performance remains remarkably stable despite the introduction of noise, with only minor degradation (less than 1%) when irrelevant retrievals are added. We also report the performance using the top-3 retrievals without explicit noise injection. The higher performance in this setting suggests that using more retrievals provides additional relevant information that the model can effectively leverage, while still filtering out any naturally occurring noise. This demonstrates the model's ability to effectively filter out irrelevant information, confirming the robustness of our approach to retrieval noise.

| Model | Entity | Query |
|---|---|---|
| Top-1 Retrieval | 20.49 | 22.19 |
| Top-1 Retrieval + 2 Noises (1) | 19.61 | 21.97 |
| Top-1 Retrieval + 2 Noises (2) | 19.63 | 22.02 |
| Top-3 Retrieval | 25.10 | 27.34 |

Table 5: Performance in accuracy (%) for RoRA-VLM with varying levels of retrieval noise on InfoSeek.

### 5.5 Domain Transfer Capability

To examine the generalizability of our approach, we conduct domain transfer experiments using the Encyclopedic-VQA dataset. The iNaturalist subset of this dataset consists of questions concerning 11 categories (e.g., Plant, Insect, Lake, etc.) of entities. To create a domain transfer setting, we select "Insect" as the target domain and modify the training set by filtering out instances from this category. We fine-tune both the baseline model and our RoRA-VLM on the original training set of the iNaturalist subset as well as the modified training set for domain transfer, and evaluate on the complete test set of the iNaturalist subset. Table 6 shows the results, where "SFT" refers to models fine-tuned on the full training set, while "Domain Transfer" refers to models fine-tuned on the modified training set (excluding "Insect" category). The results show that, even without being fine-tuned on the "Insect" category, RoRA-VLM still outperforms the baseline model that is trained on the complete training set. This demonstrates the generalizability of the cross-modal verification mechanism learned by RoRA-VLM, allowing it to effectively filter out irrelevant information even for domains not seen during training. This highlights the potential of our approach for real-world applications where domain adaptation is often required.

| Model | SFT | Domain Transfer |
|---|---|---|
| LLaVA-v1.5 | 18.23 | 17.18 |
| RoRA-VLM(ours) | 24.36 | 20.26 |

Table 6: Performance in accuracy (%) for domain transfer on Encyclopedic-VQA.

## 6 Conclusion

In this work, we introduce RoRA-VLM, a robust retrieval-augmented framework that teaches vision-language models to leverage visual information as evidence for evaluating the quality of retrieved knowledge. Unlike previous works that focus on improving retrieval quality, our work addresses the complementary challenge of how to effectively utilize retrieved knowledge. Our experimental results demonstrate that RoRA-VLM achieves strong performance on three widely adopted benchmark datasets. Through detailed ablation studies and visualizations, we demonstrate that VLMs can learn to perform cross-modal verification, mainly attending to information from retrievals containing visually similar entities to those in the query image. Furthermore, the framework shows strong generalization capabilities, including domain transfer to unseen categories, highlighting the broad applicability.

## Limitations

While our RoRA-VLM framework demonstrates significant improvements on knowledge-intensive visual question answering tasks, several limitations present opportunities for future research.

**Language Coverage** Our current evaluation is restricted to English-language datasets and benchmarks. Although our approach relies primarily on vector representations rather than natural language processing for retrieval operations, which suggests inherent compatibility with multilingual scenarios, we have not empirically validated this capability across diverse linguistic contexts. Future work should extend our evaluation framework to include multilingual knowledge-intensive VQA benchmarks to demonstrate the cross-linguistic generalizability of our vision-guided relevance assessment mechanism.

**Scale of Knowledge-Intensive Pretraining** Due to computational resource constraints, our knowledge-intensive pretraining phase is based on a subset of the WikiWeb2M dataset containing only 1 million entity-rich instances, rather than leveraging the complete Wikipedia database. While this limited-scale pretraining successfully establishes foundational visual-knowledge alignment capabilities, we anticipate that training on a more comprehensive and larger-scale knowledge repository could yield enhanced performance.

**Task Scope and Modality Extensions** Our experimental evaluation focuses exclusively on image-text visual question answering tasks, which represent only a subset of the broader vision-language domain. The underlying principles of our framework, particularly the cross-modal verification mechanism and modality-reciprocal retrieval approach, should theoretically extend to other vision-language applications. Future research directions include adapting our methodology to more vision language tasks, e.g., video understanding and three-dimensional point cloud understanding.

## References


Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *CoRR*, abs/2311.16867.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. A suite of generative tasks for multi-level multimodal webpage understanding. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. 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.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, and 24 others. 2023b. Pali-x: On scaling up a multilingual vision and language model. *CoRR*, abs/2305.18565.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, and 10 others. 2023c. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023d. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. 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.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *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*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 12031–12041. IEEE.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

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

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. *arXiv preprint arXiv:2402.08327*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thomas Mensink, Jasper R. R. Uijlings, Lluís Castrejón, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araújo, and Vittorio Ferrari. 2023. Encyclopedic VQA: visual questions about detailed properties of fine-grained categories. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3090–3101. IEEE.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. pages 8748–8763.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *CoRR*, abs/2310.01558.

## A Details of the CLIP model encoding

In this section, we provide a detailed description of how we encode an image into a sequence of visual embeddings using CLIP.

**Image Encoding with CLIP:** In the CLIP model, the visual encoder is based on the Vision Transformer (ViT) architecture. Given an image, the visual encoder processes it as a whole and encodes it into a feature representation of shape [576, 1024]. This representation can be interpreted as 576 vectors, each with a dimensionality of 1024. The 576 vectors correspond to patches of the input image, where the image is internally divided into a grid of patches during the encoding process. This division is not explicit; rather, it is an inherent part of the ViT architecture, which computes patch-level embeddings directly through a convolutional embedding layer applied to the full image. The resulting intermediate patch embeddings collectively form the image's representation in the model's latent space.

**Dimensionality of Visual Embeddings:** After passing through the vision transformer (ViT) layers, each patch is represented as a feature vector with a dimensionality of 1024. To further process these features, we utilized the final visual projection layer of the original CLIP model. This projection layer, which is also used for the pooled [CLS] token in the original implementation, is applied to all 576 patch-based feature vectors in our approach. The projection reduces the dimensionality of each feature vector from 1024 to 768. To clarify further, the visual projection layer is part of CLIP's original implementation. While it is typically applied only to the pooled [CLS] token to produce the image-level feature representation, in our work, we extend its application to all 576 patch-level feature vectors. As a result, the output is a feature representation of shape [576, 768], where 576 corresponds to the number of patches and 768 is the dimensionality of the projected patch embeddings.

After computing the patch embeddings, for each text query, we derive a 768-dimensional vector from the [CLS] token of the CLIP text encoder. We then compute the similarities between the text embedding and the image patch embeddings to select the top-m relevant patches, which are subsequently projected into the LLM's latent space using the LLaVA projector.

## B Image-Anchored Entity Retrieval

In this section, we provide a detailed explanation of the Image-Anchored Entity Retrieval component of our Multimodal-Reciprocal Retrieval method. This component uses the input query image as an anchor to retrieve visually similar images along with their associated entity information.

**Database Construction** The image database is built upon Wikipedia Image Text (WIT) dataset (Srinivasan et al., 2021), which contains 37.6 million entity-rich image-text pairs. Each text entry provides the name and background information of the entity depicted in the corresponding image, sourced from Wikipedia. To enable efficient retrieval, we encode each image in WIT into a vector using the CLIP (Radford et al., 2021) image encoder and construct a dense vector-search database based on a hierarchical navigable small-world (HNSW) graph (Malkov and Yashunin, 2018).

In this database, the encoded image features $\mathbf{z}_i = \text{CLIP}(\tilde{v}_i) \in \mathbb{R}^d$, where $d$ is the dimension of the CLIP embedding, serve as search indexes $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$. The corresponding entity names and background information for these images are stored as search values $\tilde{E} = \{\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_N\}$, where $\tilde{e}_i$ denotes the entity name and background information for candidate image $\tilde{v}_i$ and $N$ is the total number of entries in the database.

**Retrieval Process** Given a query image $v$, the image retriever $\phi^{\text{v}}$ leverages a non-parametric function to measure the cosine similarity between the CLIP embedding of the query image and all search indexes. The score of each candidate image $\tilde{v}_i$ with search index $\mathbf{z}_i$ can be expressed as:

$$P(\tilde{v}_i|v, \mathbf{Z}) = \frac{\exp\left(\text{Sim}(v, \mathbf{z}_i)\right)}{\sum_{j=1}^{N} \exp\left(\text{Sim}(v, \mathbf{z}_j)\right)}, \quad (4)$$

where the similarity function is defined as:

$$\text{Sim}(v, \mathbf{z}_i) = \frac{\text{CLIP}(v)^\top \mathbf{z}_i}{|\text{CLIP}(v)||\mathbf{z}_i|} \quad (5)$$

This function computes the cosine similarity between the CLIP embedding of the query image $v$ and the pre-computed CLIP embedding $\mathbf{z}_i$ of each candidate image in the database. Based on this similarity function, the image retriever $\phi^{\text{v}}$ fetches the top-$k$ images that are most similar to the query image along with their associated entity names and

background information:

$$(\tilde{v}_1, \tilde{e}_1), (\tilde{v}_2, \tilde{e}_2), \ldots, (\tilde{v}_k, \tilde{e}_k) = \phi^{\text{v}}(v, \mathbf{Z}, \tilde{E}), \tag{6}$$

This results in a set of top-$k$ most similar images $\tilde{V} = \{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k\}$ along with their associated entity names and background information $\tilde{E} = \{\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_k\}$.

## C Visual Token Refinement

In this section, we provide a detailed explanation of the Visual Token Refinement component of our Visual Token Refinement method. This component aims to filter out query-irrelevant visual information within both the query image and the retrieved images. This filtering process ensures that the model focuses on the most query-relevant visual features when performing cross-modal verification, enabling more accurate assessment of retrieval relevance.

**Input Query Encoding** Given a text query $t$ alongside a query image $v$, we first encode the text query using the CLIP text encoder, producing a text embedding $\mathbf{t} \in \mathbb{R}^d$, where $d$ is the embedding dimension. Similarly, the query image $v$ is encoded into a sequence of visual embeddings $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\} \in \mathbb{R}^{n \times d}$, where $\mathbf{v}_i \in \mathbb{R}^d$ denotes a visual token embedding corresponding to an image patch, and $n$ is the number of visual tokens extracted from the image by the vision encoder.

**Query Image Token Selection** For each visual token embedding $\mathbf{v}_i$ in the query image, we calculate its similarity to the text embedding by computing the dot product: $s_i = \mathbf{v}_i \cdot \mathbf{t}$. This similarity score measures how well each visual token aligns with the text query. We then select the top-$m$ visual tokens with the highest similarity scores, forming the refined visual token sequence $\hat{\mathbf{V}} \in \mathbb{R}^{m \times d}$ for the query image:

$$\hat{\mathbf{V}} = \text{Top-}m\left(\left\{\mathbf{v}_i \middle| s_i\right\}_{i=1}^{n}\right), \tag{7}$$

where $s_i = \mathbf{v}_i \cdot \mathbf{t}$ represents the similarity between visual token $\mathbf{v}_i$ and text query embedding $\mathbf{t}$.

**Retrieved Image Token Selection** Similarly, we encode each retrieved image $\tilde{v}_i \in \tilde{V}$ into a sequence of visual token embeddings $\tilde{\mathbf{V}}_i = \{\tilde{\mathbf{v}}_{i,1}, \tilde{\mathbf{v}}_{i,2}, ..., \tilde{\mathbf{v}}_{i,n}\} \in \mathbb{R}^{n \times d}$. For each visual token embedding $\tilde{\mathbf{v}}_{i,j} \in \mathbb{R}^d$ in the retrieved image, we compute its similarity to the refined query image tokens by calculating the sum of its dot product with all of the selected visual tokens from the query image:

$$s_j = \sum_{i=1}^{m} (\mathbf{v}_i \cdot \tilde{\mathbf{v}}_{i,j}) \tag{8}$$

where $\mathbf{v}_i \in \hat{\mathbf{V}}$. We then select the top-$m$ most relevant visual tokens of the retrieved image, forming the refined visual token sequence $\hat{\mathbf{V}}_i \in \mathbb{R}^{m \times d}$ for each retrieved image:

$$\hat{\mathbf{V}}_i = \text{Top-}m\left(\left\{\tilde{\mathbf{v}}_{i,j} \middle| \sum_{i=1}^{m} s_j)\right\}_{j=1}^{n}\right). \tag{9}$$

**Details of the pooling process** In our implementation, each image is processed into a feature matrix with shape [576, 768] by the CLIP visual encoder and the LLaVA projector. Our Visual Token Refinement method selects the top 144 visual tokens that are most relevant to the query, constructing a feature matrix of shape [144, 768]. This selection process enables the VLM to focus more effectively on query-relevant image content while mitigating the influence of irrelevant noise, such as image backgrounds or query-irrelevant entities present in the image. For comparison purposes in ablation studies, we implemented an average-pooling-based baseline that processes the same [576, 768] visual patch vectors into [144, 768] vectors. Specifically, we reshape the 576 patch vectors into a 24 × 24 grid corresponding to the spatial arrangement of patches in the original image, then apply a 2D average pooling operation with a kernel size of 2 × 2 and a stride of 2. This pooling reduces the spatial resolution from 24 × 24 to 12 × 12, yielding 144 patch vectors while maintaining the 768-dimensional feature vector for each patch. By reducing the number of feature vectors from 576 to 144, this process ensures compatibility with the limited sequence length of the LLM and aligns the number of input tokens for both methods, allowing for direct and fair comparison in the ablation studies.

### C.1 Robustness of RoRA-VLM Under Varying Levels of Retrieval Noise

To further analyze the ability of our RoRA-VLM to handle noisy retrieval and validate its robustness, we conducted additional ablation studies involving controlled retrieval noise scenarios. The key challenge in ideally proving the effectiveness of our model in ignoring retrieval noise is the lack of

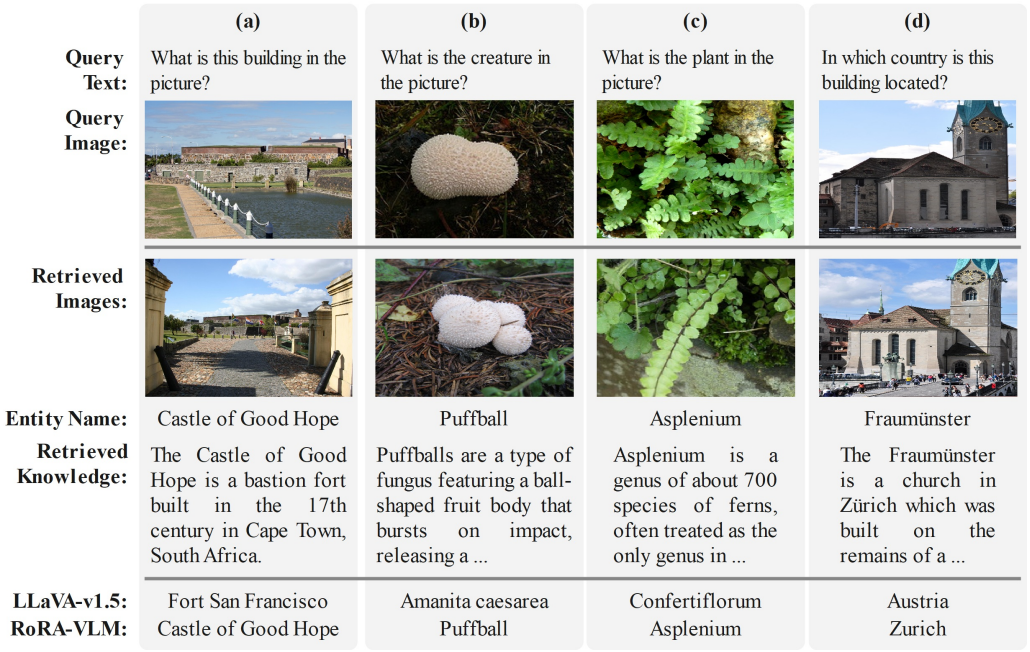


Figure 5: Qualitative results of the Multimodal-Reciprocal Retrieval.

gold-standard labels for the retrieval process in the evaluation datasets. Specifically, we do not have precise relevancy labels between input queries and all candidate samples for retrieval, making it infeasible to construct an experiment with exactly one relevant sample and two randomly sampled irrelevant samples. Therefore, we designed an alternative experiment with varying levels of retrieval noise. During the inference stage, instead of using the top-3 retrieved entity images and their corresponding knowledge snippets, we tested a setting where we used the top-1 retrieved entity image and its knowledge snippet along with two randomly sampled irrelevant entity images and their knowledge snippets. This random sampling process was repeated twice, resulting in two distinct sets of irrelevant entity images and knowledge snippets for the same input instance. Additionally, we tested another setting using only the top-1 retrieved entity image and its corresponding knowledge snippet for generation augmentation. Using these four configurations of retrieved entity images and knowledge snippets, we evaluated retrieval augmentation on the InfoSeek dataset. The results are summarized in the Table 5.

## D Evaluation of the Multimodal-Reciprocal Retrieval

Figure 5 presents several examples for qualitative analysis. Our retrieval method effectively identifies images that contain entities matching those in the query images. Although the perspectives of the entities in the retrieved images differ from those in the query images, the retrieved images provide sufficient visual attributes for entity identification (e.g., the gap in the wall in Figure 5(a) and the shape of the leaves in Figure 5(c)).

| Model | Entity | Query |
|---|---|---|
| LLaVA-1.5 | 10.34 | 12.98 |
| RoRA LLaVA-1.5 | **25.10** | **27.34** |
| - w/o WikiWeb2M | 20.68 | 23.41 |
| - w/ ShareGPT4V | 21.28 | 22.84 |

Table 7: Ablation studies of different pre-training configurations. Performance is reported in accuracy (%) on InfoSeek.

## E Ablation Studies

### E.1 Effect of Knowledge-Intensive Pre-training

To assess the impact of knowledge-intensive pre-training on cross-modal verification capabilities, we conducted two experiments with different pre-training configurations and reported the results in Table 7. Directly fine-tuning without the knowledge-intensive pre-training on WikiWeb2M downgrades the RoRA LLaVA-1.5 model performance from 25.10% to 20.68% on InfoSeek-Entity

Table 8: Performance comparison in accuracy (%) for VLMs with different numbers of retrieval knowledge snippets on the InfoSeek.

| Model | Entity | Query |
|---|---|---|
| LLaVA-v1.5 | | |
| - 4 snippets | 20.68 | 23.41 |
| - 8 snippets | 20.84 | 23.34 |
| RoRA-VLM(ours) | | |
| - 4 snippets | 24.56 | 26.33 |
| - 8 snippets | 25.10 | 27.34 |

and from 27.34% to 23.41% on InfoSeek-Query. We also compared pre-training on WikiWeb2M with pre-training on ShareGPT4V, a generic image-caption dataset where captions primarily describe image content without detailed entity information. The results show that RoRA-VLM pre-trained on WikiWeb2M outperforms the same model pre-trained on ShareGPT4V by 3.28% on InfoSeek-Entity and 4.5% on InfoSeek-Query. This highlights the importance of alignment between the visual appearances and entity knowledge for knowledge-intensive tasks.

### E.2 Effect of the Number of Retrieved Knowledge Snippets

We investigate the impact of the number of textual knowledge snippets $l$ returned for each image during the query-expanded text retrieval, and show the results on the InfoSeek dataset in Table 8. LLaVA-v1.5 with 4 or 8 snippets denotes the LLaVA-v1.5 fine-tuned with retrieval augmentation but without visual token refinement and knowledge-intensive pertaining. As shown in the table, expanding the retrieval from top-4 to top-8 snippets results in marginal improvements, demonstrating the less sensitivity of our multimodal-reciprocal retrieval on the number of retrieved knowledge snippets.

### E.3 Effect of Truncation

We implement a truncation strategy for each retrieved knowledge snippet during tokenization to construct the multimodal interleaved input, preventing longer preceding retrieved knowledge snippets from dominating the limited input sequence space, thereby ensuring that subsequent retrieved information is preserved. However, this raises an important question: how much valuable information is lost due to this truncation?

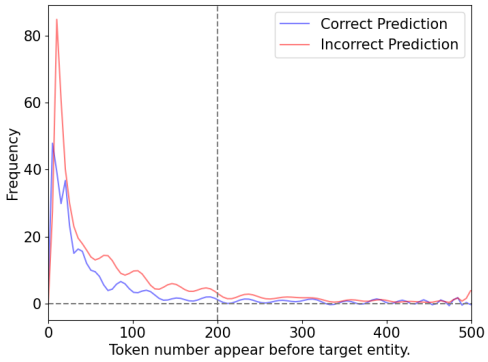


Figure 6: Position distribution of the target entity name within retrieved knowledge snippets.

To assess the potential loss of critical information, we examine instances where the retrieved knowledge snippets explicitly mention the target entity name. We count the number of tokens that appear before this mention and visualize the positional distribution of key information (i.e., the target entity name) within the retrieved snippets, as shown in Figure 6. As depicted, in most cases, the entity name appears within the first 200 tokens of the retrieved passages, whereas our truncation is applied at the 400-token mark for each passage. This buffer ensures a high retention rate of valuable information, minimizing the risk of discarding critical content due to truncation.

## F Experiment Setup Details

### F.1 Datasets

**OVEN (Hu et al., 2023)** OVEN is an entity recognition dataset constructed by repurposing 14 existing datasets, comprising over 5 million instances. All labels in OVEN are mapped onto a unified label space of Wikipedia entities. Each instance consists of an entity image paired with its corresponding entity name. The tasks in OVEN require vision-language models (VLMs) to recognize visual entities from a pool of six million possible Wikipedia entities.

**InfoSeek (Chen et al., 2023d)** InfoSeek is a large-scale visual question answering (VQA) dataset focused on knowledge-seeking queries. It consists of over 1.35 million image-text pairs, each posing various questions about objects, scenes, and actions that require external knowledge—such as factual information—rather than solely relying on the visual content.

**Encyclopedic-VQA (Mensink et al., 2023)** Encyclopedic-VQA is a knowledge-intensive VQA dataset containing over 221,000 image-text instances that require deep reasoning and access to external knowledge. It is well-suited for evaluating a model's ability to answer questions that extend beyond the image content.

### F.2 Baselines

**Baselines** We compare our framework with several state-of-the-art vision-language models. **LLaVA-v1.5** (Liu et al., 2023a) integrates pre-trained visual and language models for strong performance in multimodal tasks, while **LLaVA-v1.6** (Liu et al., 2024) introduces improved fine-tuning techniques. **PaLI-17B** (Chen et al., 2023c) utilizes a 17-billion-parameter architecture, excelling in image captioning and visual question answering, with **PaLI-X** (Chen et al., 2023b) improving performance on vision-language tasks by scaling up the model size and incorporating a high-capacity visual encoder. **BLIP-2** (Li et al., 2023) introduces efficient visual grounding through a Q-former, and **InstructBLIP** (Dai et al., 2023) enhances it for instruction-following tasks. **CLIP2CLIP** (Hu et al., 2023) leverages a CLIP-based model for improved image captioning. Recent work **Wiki-LLaVA** (Caffagni et al., 2024) is designed for entity-centric question answering, aligning visual data with external knowledge from Wikipedia. **PreFLMR** (Lin et al., 2024) introduces a robust multimodal retriever pre-trained on a vision-language corpus comprising over ten million samples, enabling high-quality retrieval to augment the generation processes. **RA-CM3** (Yasunaga et al., 2023) employs a cross-modality retrieval mechanism to access and leverage multimodal information to enhance the performance of multimodal generation. To ensure a fair comparison, all the baseline models are fine-tuned on the OVEN (Hu et al., 2023), InfoSeek (Chen et al., 2023d), and Enc-VQA (Mensink et al., 2023) datasets respectively, and then evaluated on the corresponding tasks.

### F.3 Implementation Details

We adopt `LLaVA-v1.5-7B` (Liu et al., 2023a) as the backbone model for our RoRA-VLM. In our experiments, limited by the input sequence length, we set the retrieval parameters as follows: $k = 3$ and $l = 3$ for multimodal-reciprocal retrieval, and $m = 144$ for our visual token refinement method.

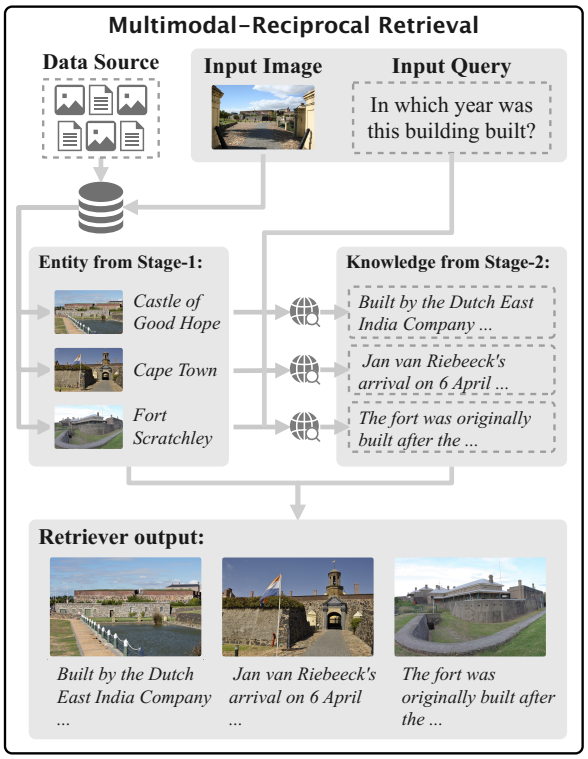


Figure 7: Overview of the Multimodal-Reciprocal Retrieval

All models are trained using 8 NVIDIA H100 GPUs. Both pre-training and fine-tuning processes follow the hyperparameters specified in the original LLaVA (Liu et al., 2023a) setup, ensuring consistency with previous work.

### F.4 Schematic Diagram of the Multimodal-Reciprocal Retrieval

We include Figure 7 to provide a more intuitive explanation of our proposed Image-Anchored Entity Retrieval and Query-Expanded Text Retrieval.

### F.5 Schematic Diagram of the Visual Token Refinement

We include Figure 8 to provide a more intuitive explanation of the visual token refinement .

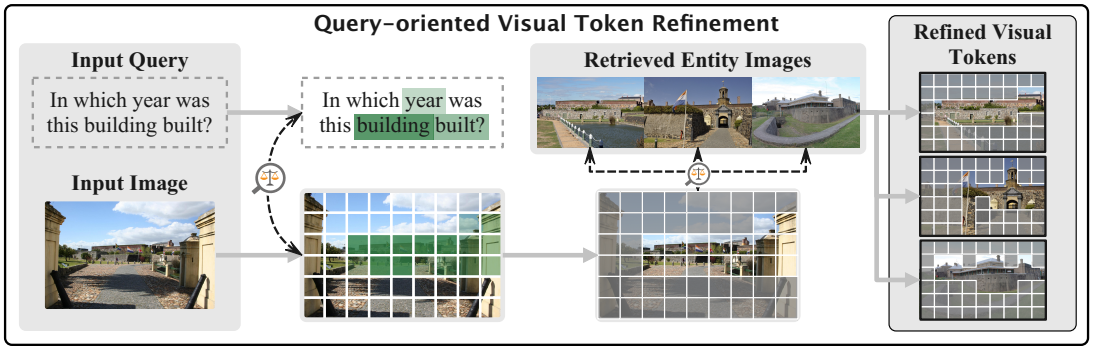


Figure 8: Overview of the Visual Token Refinement