VAMER: Visual-Anchored Multimodal Evidence Reasoning for Knowledge-Based VQA

Jiuxiang You , Zhenguo Yang * , Xiaoping Li , Qing Li , Yi Yu *

Posted Date: 25 May 2026

doi: 10.20944/preprints202605.1648.v1

Keywords: knowledge-based VQA; visual entity linking; multimodal evidence chain

 doi:10.20944/preprints202605.1648.v1

Article

VAMER: Visual-Anchored Multimodal Evidence Reasoning for Knowledge-Based VQA

Jiuxiang You [1], Zhenguo Yang [2,*], Xiaoping Li [2], Qing Li [3] and Yi Yu [1,4,*]

[1] Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima, Japan
[2] School of Computer Science, Guangdong University of Technology, Guangzhou, China
[3] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[4] Education and Research Center for Artificial Intelligence and Data Innovation, Hiroshima University, Hiroshima, Japan
* Correspondence: yzg@gdut.edu.cn (Z.Y.); yiyu@hiroshima-u.ac.jp (Y.Y.)

Abstract

Knowledge-Based Visual Question Answering (KB-VQA) relies on external knowledge for cross-modal scene understanding and reasoning. Existing methods still suffer from limited reasoning capability due to two major drawbacks: (1) the visual entity anchoring issue, where current methods fail to accurately anchor visual entities from questions, leading to irrelevant knowledge retrieval and misleading reasoning. (2) the visual-aware reasoning issue, where prior approaches overly rely on text-only reasoning while ignoring visual cues, resulting in unreliable reasoning chains. To this end, we propose VAMER, a Visual-Anchored Multimodal Evidence Reasoning framework with two components: (1) For the visual entity anchoring issue, we introduce a Visual Entity Linking (VEL) module that utilizes the reasoning capability of a Visual-Language Model (VLM) to extract semantic and spatial information from questions, which is used to guide semantic-spatial contrastive learning for entity localization. (2) For the visual-aware reasoning issue, we propose a Multimodal Evidence Chain Reasoning (MECR) module that adopts a hierarchical two-phase approach to separately handle evidence chain construction and answer generation, enabling iterative integration of visual and textual information for improved reasoning reliability. Extensive experiments on the OK-VQA, A-OKVQA, and F-VQA datasets demonstrate the effectiveness of the proposed method for Knowledge-based VQA.

Keywords: knowledge-based VQA; visual entity linking; multimodal evidence chain

1. Introduction

With the rapid development of Artificial Intelligence (AI) and computer vision, visual content understanding has become crucial in human-machine interaction. In this context, Visual Question Answering (VQA)[1] emerges as a fundamental challenge that requires deep comprehension of both visual and textual information [2–8]. However, visual content alone is insufficient for complex reasoning scenarios that demand common sense and real-world knowledge. For instance, to answer questions about brand relationships, historical contexts, or functional properties, systems need to go beyond pixel-level understanding. This demand has motivated the development of Knowledge-Based Visual Question Answering (KB-VQA) [9–11], which integrates visual perception with knowledge-driven reasoning. In the realm of KB-VQA, both visual understanding and knowledge reasoning are essential for comprehensive scene interpretation. Visual understanding helps locate and identify specific entities in complex scenes, while knowledge reasoning enables inference based on external information. As shown in Figure 1, when answering the question "What brand made the shoes the hatless man on the left has on?", the system faces two fundamental challenges: (1) Accurately identifying and locating the specific shoes being referenced, especially when multiple people and shoes are present; (2) Connecting the visual evidence with relevant brand knowledge to make a reliable inference. These challenges are particularly acute in real-world applications where questions usually contain complex spatial

 doi:10.20944/preprints202605.1648.v1

references and require multi-step reasoning. To address these challenges, prior approaches have adopted three main paradigms: explicit methods, implicit methods, and hybrid approaches. Explicit methods [12–16] primarily rely on structured knowledge bases, evolving from simple entity matching to sophisticated semantic parsing approaches. For instance, Gao et al. [17] propose to transform visual content into textual descriptions through object detection and OCR for knowledge retrieval. However, these approaches usually fail to handle complex spatial references and ambiguous entity mentions, leading to irrelevant knowledge retrieval, as shown in Figure 1(b). Implicit methods leverage large language models [18] and multimodal models [19] to capture semantic relationships, demonstrating impressive results on simple queries but falling short on questions requiring precise visual grounding and structured reasoning. Recent hybrid approaches [20–22] attempt to combine both paradigms, yet they still face two critical limitations: (1) insufficient visual entity anchoring and (2) inadequate visual-aware reasoning resulting in unreliable inference chains. As exemplified in Figure 1(a), MM-COT [22] relies solely on text-based reasoning, neglecting visual cues and causing factual inconsistencies in its generated rationales, despite having access to relevant knowledge.

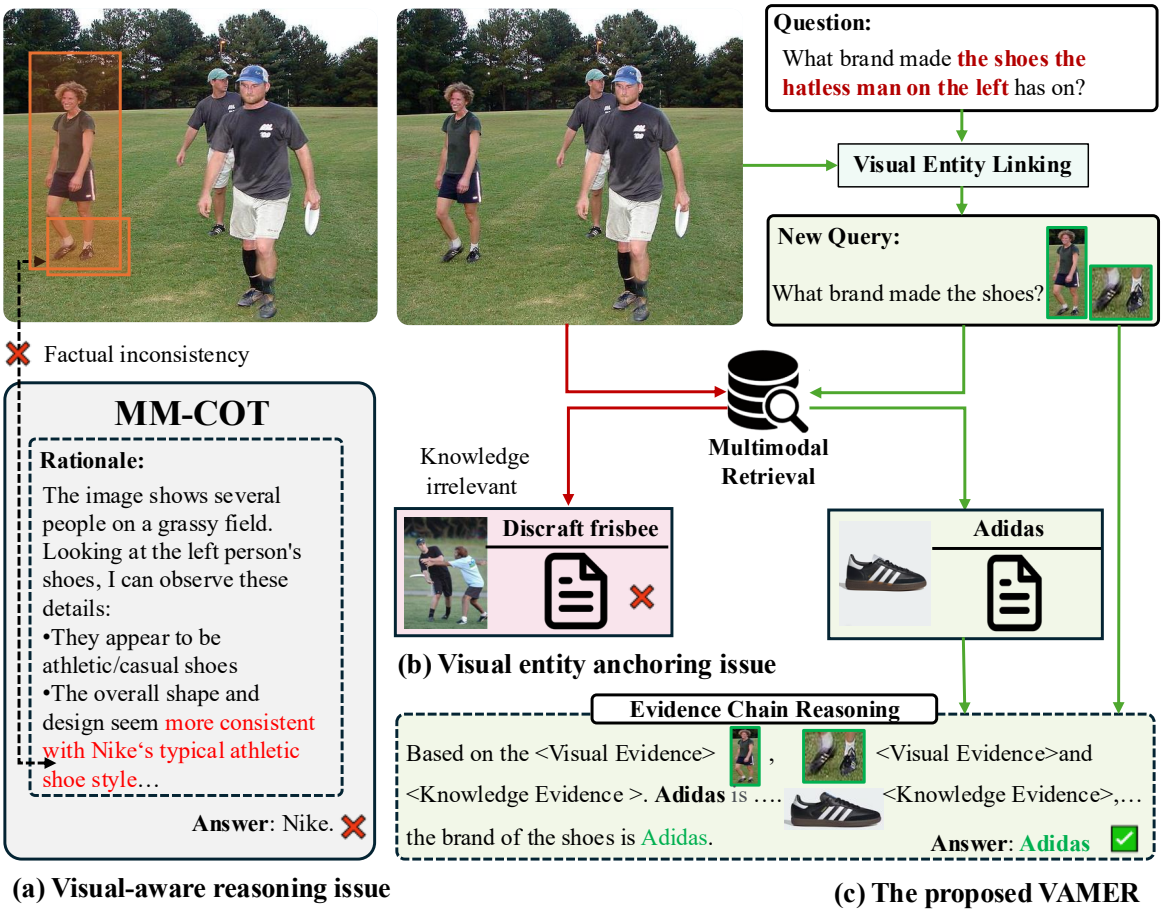


Figure 1. Example of visual-anchored multimodal evidence reasoning for KB-VQA.

To this end, we propose VAMER to construct reliable reasoning chains through visual-anchored multimodal evidence. The framework consists of two components: (1) Visual Entity Linking (VEL) module, which leverages vision-language model-driven semantic cues to guide semantic-spatial contrastive learning for localization of key visual entity regions through visual-linguistic alignment; (2) Multimodal Evidence Chain Reasoning module (MECR), which employs a hierarchical fine-tuning strategy to decouple evidence generation from answer inference, ensuring reliable reasoning paths grounded in both visual and knowledge context. As a result, VAMER can handle complex queries that are challenging for previous methods. For instance, as shown in Figure 1(c), it successfully

 doi:10.20944/preprints202605.1648.v1

identifies specific shoes and infers their brand through reliable evidence chains. In summary, the main contributions of this work are as follows:

- We propose a novel Visual-Anchored Multimodal Evidence Reasoning (VAMER) framework that effectively addresses the challenges of visual entity anchoring and visual-aware reasoning in KB-VQA through VLM-driven entity localization and hierarchical knowledge reasoning.
- We design a VEL module that utilizes VLM-driven semantic cues to guide semantic-spatial contrastive learning, aligning visual-linguistic and spatial features for key visual entity localization.
- We present a multimodal evidence chain reasoning module with a hierarchical fine-tuning strategy that decouples evidence chain generation from answer inference to improve both answer accuracy and reasoning reliability.
- We conduct extensive experiments on OK-VQA, A-OKVQA, and F-VQA datasets to demonstrate that the proposed VAMER achieves outstanding performance and analyze the effectiveness of each module.

The remainder of the paper is organized as follows. Section II reviews related work in knowledge-based VQA, visual entity linking, and multimodal reasoning. Section III describes the detailed architecture and methodology of VAMER. Section IV provides comprehensive experimental results and analysis. Section V concludes the paper, followed by acknowledgements.

2. Related Work

This section reviews previous studies on visual question answering and related techniques, with a focus on knowledge-based VQA, visual entity linking, and multimodal evidence chain reasoning.

2.1. Knowledge-Based VQA

Knowledge-based VQA [9–11] extends traditional VQA to scenarios requiring commonsense knowledge. Existing KB-VQA methods can be grouped into three categories: explicit, implicit, and hybrid methods. Explicit methods utilize textual knowledge sources, such as Wikipedia [23] and ConceptNet [24]. For instance, Gao et al. [17] adopt techniques like object detection [25], captioning [26], and Optical Character Recognition (OCR) to convert visual content into text and combine it with questions to retrieve knowledge from Wikipedia. Nonetheless, queries involving multiple entities inevitably introduce irrelevant knowledge during retrieval. Implicit knowledge methods aim to bridge gaps in semantic relationships and commonsense understanding. For example, Yang et al. [18] integrate image captioning with knowledge reasoning by prompting GPT-3 [18]. Rather than substituting images with captions, multimodal large-scale models such as Flamingo [19] can directly infer answers from original image-question pairs. However, these models may struggle with fine-grained knowledge queries, especially when images lack adequate information. Some KB-VQA methods incorporate both implicit and explicit knowledge into visual reasoning. For instance, Gui et al. [27] propose a framework that leverages both Wikipedia and GPT-3 for integrated reasoning. Recently, Khademi et al. introduced MM-Reasoner [28], which integrates outputs from multiple vision APIs, including dense captioning, object detection, and OCR, to extract image information. This information is then processed by large language models (LLMs) for reasoning and by vision-language models (VLMs) for prediction. Despite these advances, using multiple independent APIs can cause information overload and lack explicit anchoring for visual entities, generating noise and resulting in erroneous reasoning.

2.2. Visual Entity Linking

Entity linking aligns textual mentions with specific entities in a knowledge graph and was originally developed for textual data. The rise of multimodal content, such as image-text pairs, introduces the challenge of incorporating visual information. Karpathy et al. [29] align textual mentions with entities in image regions. As social media rapidly evolves, entity linking in text cannot handle multimodal posts, such as paired images and text. Zhu et al. [30] propose a cross-modal

 doi:10.20944/preprints202605.1648.v1

alignment network to link visual regions with text mentions. Venkitasubramanian et al. [31] align visual mentions with entities in knowledge graphs. Moon et al. [32] introduce Multimodal Name Entity Disambiguation (MNED) to disambiguate entities in both text and images. Zheng et al. [33] combine visual and textual features with knowledge graph data for linking object regions to entities. Gan et al. [34] propose a bipartite graph matching approach for entity linking and introduce the Multimodal Movie Entity Linking (M3EL) task and dataset. Many existing methods effectively link text mentions or text-image pairs to knowledge graphs but neglect ambiguous visual references in images, which are important in question-answering contexts.

2.3. Multimodal Reasoning

Visual reasoning requires both perceptual and high-level cognitive abilities [35], particularly for tasks demanding precise entity understanding and complex inference. Several tasks evaluate visual reasoning capabilities, including Visual Question Answering (VQA) [36] and Visual Entailment [37]. Traditional vision-language models adopt neuro-symbolic approaches [38,39] to explicitly model the reasoning process. However, such methods typically struggle with complex scenarios requiring accurate entity localization and knowledge grounding. The emergence of Large Language Models (LLMs) has brought new perspectives to visual reasoning. Recent vision-language models have started to leverage LLMs' advanced reasoning abilities for visual tasks. Various approaches have been proposed, including optimizing visual encoding strategies [40] for better visual understanding, positioning LLM as a decision-making agent [41] with task-specific visual modules, and improving reasoning through sequential instruction tuning. However, these methods rely on text-based reasoning during inference, overlooking essential visual details and failing to capture fine-grained visual cues critical for accurate reasoning.

3. Methodology

In this section, we introduce the framework of VAMER in detail. Section III.A describes the overview of the framework. Section III.B describes the architecture of the Visual Entity Linking (VEL) module. Section III.C describes Multimodal Knowledge Retriever (MKR) module. Section III.D presents the Multimodal Evidence Chain Reasoning module (MECR).

3.1. Overview of the Framework

Given a question (Q) and a visual image (I), we propose a Visual-Anchored Multimodal Evidence Reasoning framework (VAMER) as shown in Figure 2, which consists of a chain of evidence construction stage and a chain of evidence reasoning stage. In the first stage, the framework constructs multimodal queries $\widetilde{Q}$ through LLM-driven Semantic and Spatial Contrastive Learning (CL), generating new queries for knowledge retrieval from a multimodal knowledge base. In the second stage, the framework integrates visual evidence, knowledge evidence, and queries through a visual-language model [42] to generate explanations and final answer. The evidence chain-based approach enables effective reasoning by combining visual information, external knowledge, and question understanding for knowledge-based VQA.

 doi:10.20944/preprints202605.1648.v1

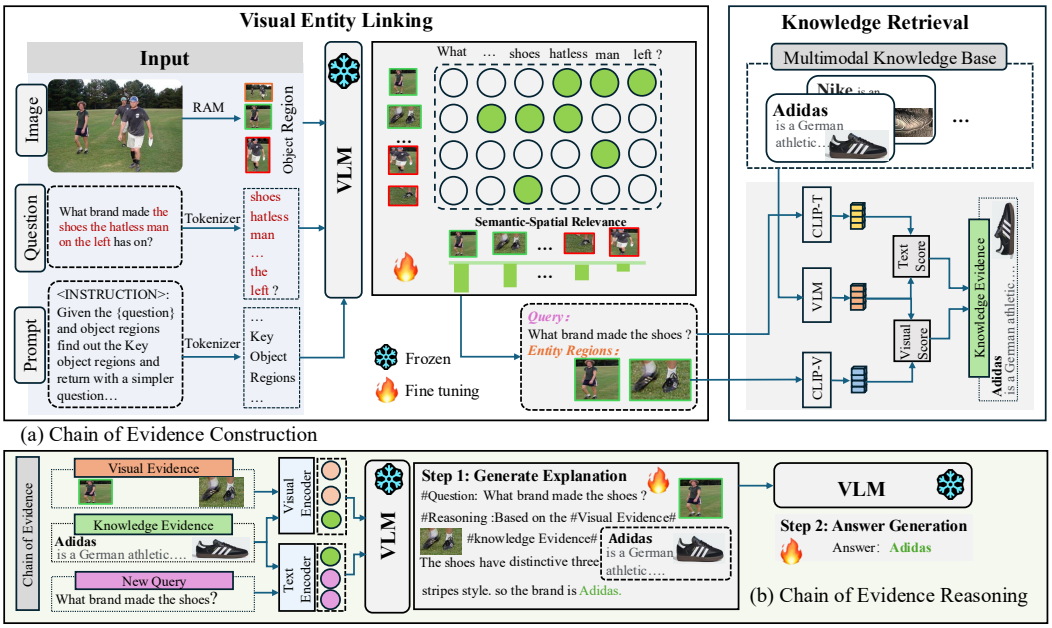


Figure 2. Given an image-question pair (I, Q), the framework consists of two stages: (1) chain of evidence construction, where VEL identifies relevant visual entities to form a refined query $\widetilde{Q}$ for knowledge retrieval (MKR); (2) chain of evidence reasoning, where visual evidence, retrieved knowledge, and refined query are combined by MECR to generate a reasoning chain R and the answer A.

3.2. Visual Entity Linking Module

Due to ambiguities in spatial and semantic alignment, existing methods [12–14] struggle to accurately link visual entities in questions to their corresponding image regions. In this section, we propose the Visual Entity Linking (VEL) module, which aims to link key visual entities mentioned in the questions to generate a multimodal query combining entity regions and refined question text. Given a question Q and an image I, our goal is to identify the most relevant regions that align with the question description. As shown in Figure 3, the framework consists of four main stages:

1) Question Parsing and Feature Extraction. Given an input image I and a question Q, we first extract region proposals using RAM [43]:

$$R = \{r_i\}_{i=1}^{N_r} = RAM(I) \tag{1}$$

where N_r is the number of candidate regions. Meanwhile, a VLM [42] parses the question Q into semantic entities and spatial relationships:

$$T_{sem},\ T_{spa} = VLM(Q, I, \text{Prompt 1}) \tag{2}$$

2) Multi-modal Feature Processing. Each candidate region, the semantic phrase, and the spatial phrase are projected into a common d_L-dimensional space:

$$f_{r_i} = \text{CLIP}(r_i) \in \mathbb{R}^{d_L} \tag{3}$$

$$f_{sem} = \text{VLM}(T_{sem}) \in \mathbb{R}^{d_L} \tag{4}$$

$$f_{spa} = \text{VLM}(T_{spa}) \in \mathbb{R}^{d_L} \tag{5}$$

where f_{r_i} denotes the visual feature of region r_i, while f_{sem} and f_{spa} are textual features for semantics and spatial relations, respectively.

3) Dual Contrastive Learning. Using the semantic entities and spatial relationships extracted by Prompt 1, along with the object regions identified by Prompt 2, the VEL module performs dual contrastive learning to identify and link relevant regions while simultaneously generating a refined query. Specifically, we employ parallel contrastive objectives: a semantic contrastive learning (Semantic

 doi:10.20944/preprints202605.1648.v1

CL) that aligns region features with text of semantic entities, and spatial contrastive learning (Spatial CL) that captures positional dependencies. The contrastive loss in Semantic CL is defined as follows:

$$\mathcal{L}_{\text{sem}} = -\log \frac{\exp(\text{sim}(f_{sem}, f_{r_{i^+}})/\tau)}{\sum_{i=1}^{N_r} \exp(\text{sim}(f_{sem}, f_{r_i})/\tau)} \tag{6}$$

Spatial CL is defined as follows:

$$\mathcal{L}_{\text{spa}} = -\log \frac{\exp(\text{sim}(f_{spa}, f_{r_{i^*}})/\tau)}{\sum_{i=1}^{N_r} \exp(\text{sim}(f_{spa}, f_{r_i})/\tau)} \tag{7}$$

where $f_{r_{i^+}}$ and $f_{r_{i^*}}$ denote the positive region features for the semantic and spatial branches, respectively. $\text{sim}(a, b) = \frac{a^\top b}{\|a\|\|b\|}$ is cosine similarity and τ is the temperature.

Prompt Template 1

`/* Instruction */`
Task: *Extract semantic entity and spatial relationship*
Input: [question]
Please identify:

1. Semantic entity: The core object/entity
2. Spatial relationship: The location reference

Output format:

- Semantic entity: [entity description]
- Spatial relationship: [spatial reference]

`/* Instruction */`

Prompt Template 2

`/* Instruction */`
Given: [question] and [object_regions]
Tasks:

1. *Identify key object regions*
2. *Generate a simpler question*

Output format:

- Key regions: [region_ids]
- Refined question: [new_question]

`/* Instruction */`

Through the training, in the Semantic CL branch, the features from the semantic phrase are pulled closer to their corresponding positive object region samples while being pushed away from negative region samples in the batch. Similarly, in the Spatial CL branch, the features from the spatial phrase learn to discriminate between correct and incorrect region samples based on their spatial positions. In this way, the dual contrastive learning helps capture both object-text semantic correspondence and spatial-region dependencies. The overall loss can be expressed as below:

$$\mathcal{L}_{\text{VEL}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{spa}} \tag{8}$$

4) Region scoring and multimodal query formation. To refine the initial query, each region r_i is scored by combining semantic and spatial similarity:

$$\sigma_i = \alpha \, \text{sim}(f_{sem}, f_{r_i}) + (1 - \alpha) \, \text{sim}(f_{\text{spa}}, f_{r_i}) \tag{9}$$

 doi:10.20944/preprints202605.1648.v1

where $\alpha \in [0,1]$ is a weighting factor that controls the contribution of semantic features f_{sem} and spatial features f_{spa}. We keep the top-k regions $\mathcal{S} = \text{top-k}(\{\sigma_i\})$ and ask the VLM [42] (via `Prompt 2`) to rewrite Q into a simpler text Q' focused on these regions. Let C_{r_i} be the crop of region r_i. The final refined multimodal query, to be passed to the knowledge retriever, is defined as:

$$\widetilde{Q} = \left(Q',\ \{ C_{r_i} \}_{i \in \mathcal{S}} \right) \tag{10}$$

It contains one refined text string and k key visual regions, so later modules can seamlessly perform text–image joint retrieval.

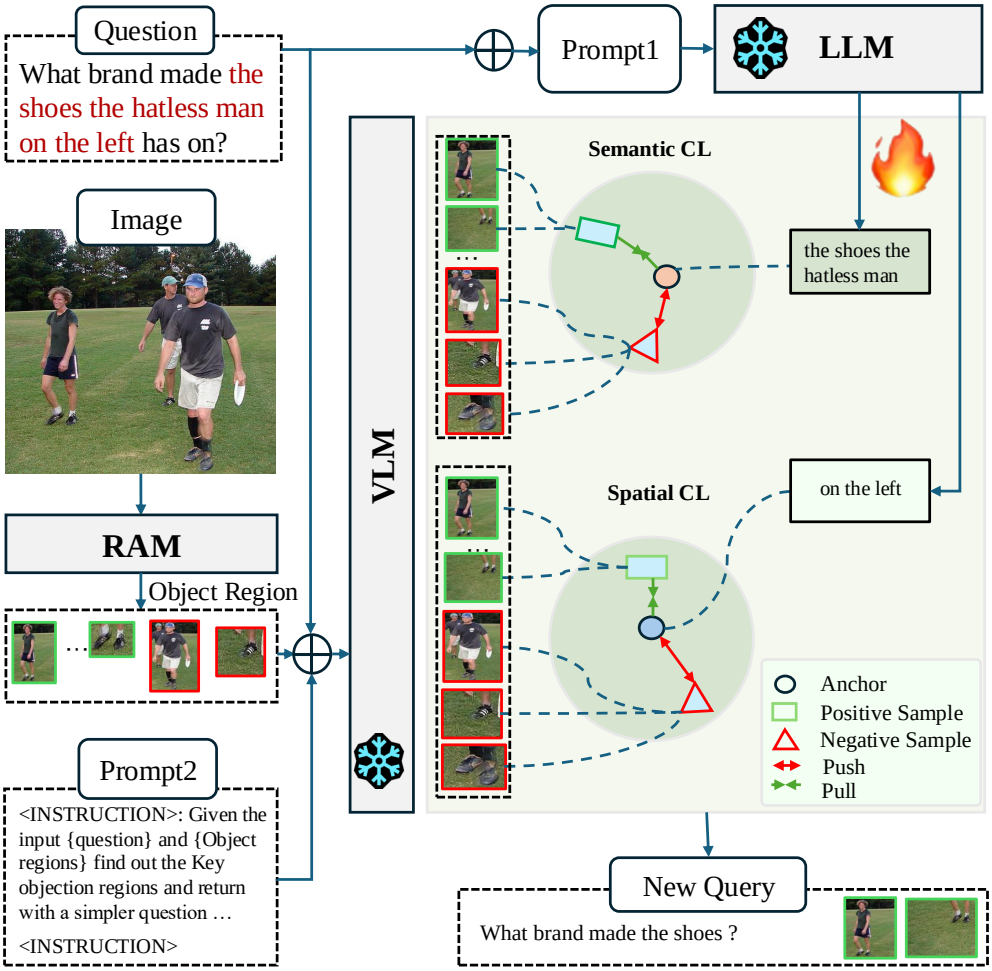


Figure 3. The structure of VEL.

3.3. Multimodal Knowledge Retriever

Given a refined multimodal query $\widetilde{Q} = (Q', \{C_{r_i}\}_{i \in \mathcal{S}})$ produced by VEL, the MKR module utilizes the query text and the visual entity regions for knowledge retrieval to compensate for the lack of textual knowledge.

1) Knowledge base format. Each entry in the knowledge base is a triple:

$$\text{kg}_j = (\text{kg}_{n_j},\ \text{kg}_{v_j},\ \text{kg}_{d_j}), \qquad j = 1, \ldots, N_{\text{kg}}, \tag{11}$$

where kg_{n_j} denotes the title, kg_{v_j} the visual content, and kg_{d_j} the textual description. N_{kg} is the total number of entries in the knowledge base.

2) Query encoding. *Visual branch.* To eliminate cross-entity interference during retrieval, we independently encode the image and each extracted key visual entity. The image I and the Regions of

 doi:10.20944/preprints202605.1648.v1

Interest (RoIs) $\{C_{r_i}\}_{i\in\mathcal{S}}$ are encoded by a CLIP-based visual encoder [44] $\mathcal{H}_v(\cdot)$, from which we take the [CLS] output so that each region yields a single d_L-dimensional vector:

$$E_v = \left[\, \mathcal{H}_v(I);\ \mathcal{H}_v(C_{r_{i_1}});\ \ldots;\ \mathcal{H}_v(C_{r_{i_k}}) \,\right] \in \mathbb{R}^{(1+k)\times d_L}. \tag{12}$$

Text branch. The refined text Q' is encoded at the token level by the CLIP text encoder [44] $\mathcal{H}_t(\cdot)$:

$$E_q = \mathcal{H}_t(Q') \in \mathbb{R}^{T_q\times d_L}, \tag{13}$$

where T_q is the number of tokens in Q'. Since $\mathcal{H}_v$ and $\mathcal{H}_t$ share the CLIP joint embedding space, E_q and E_v are directly comparable. The complete query representation is obtained by row-wise concatenation:

$$E_{\widetilde{Q}} = \left[\, E_q\,;\, E_v \,\right] \in \mathbb{R}^{N_Q\times d_L}, \quad N_Q = T_q + 1 + k. \tag{14}$$

3) Knowledge encoding. Each knowledge triple is linearised as [CLS] kg_{n_j} [SEP] kg_{d_j} [SEP] and encoded by a multilingual text encoder [45] $\mathcal{H}_m(\cdot)$:

$$E_{\mathrm{kg}_j} = \mathcal{H}_m(\mathrm{kg}_j) \in \mathbb{R}^{L_j\times d_L}, \tag{15}$$

where L_j is the number of tokens in the j-th linearised triple. A lightweight projection layer, trained jointly with the retriever, maps $\mathcal{H}_m$'s output into the CLIP embedding space used by $E_{\widetilde{Q}}$, so that all query and key vectors reside in a common d_L-dimensional space.

4) Cross-modal relevance scoring. We adopt a token-level sum-of-max similarity for cross-modal alignment:

$$\mathrm{Score}\big(\widetilde{Q},\, \mathrm{kg}_k\big) = \sum_{i=1}^{N_Q} \max_{1\le \ell\le L_k} \big[E_{\widetilde{Q}}\big]_i^{\top} \big[E_{\mathrm{kg}_k}\big]_\ell, \tag{16}$$

where $\big[\,\cdot\,\big]_i$ denotes the i-th row of a matrix, and L_k is the token length of the k-th knowledge triple. Unlike Dense Passage Retrieval (DPR) [46], which compresses $E_{\widetilde{Q}}$ and E_{kg_k} into a single dense vector representation, Eq. (16) preserves fine-grained alignment at the word, phrase, and visual-entity levels. Through independent encoding and fine-grained similarity computation, this design prevents interference between different key visual entities and textual information, benefiting the multimodal knowledge retriever.

3.4. Multimodal Evidence Chain Reasoning

In order to address complex reasoning tasks that require integrating visual and textual information, we propose a Multimodal Evidence Chain Reasoning (MECR) module to construct a coherent evidence chain. Specifically, we define the evidence components as follows:

- $\mathcal{V} = \{C_{r_i}\}_{i\in\mathcal{S}}$ – the k visual crops selected by VEL;
- $\mathcal{K} = \{\mathrm{kg}_m\}_{m\in\mathcal{M}}$ – the top-M knowledge triples returned by MKR;

These elements are aggregated into a unified evidence set:

$$\mathcal{E} = \{\mathcal{V}, \mathcal{K}\} \tag{17}$$

which serves as the input to the subsequent reasoning stages.

Stage 1: Chain-of-Evidence Reasoning

Given the evidence set $\mathcal{E}$ and the refined query $\widetilde{Q}$, we fine-tune the VLM [42] to generate a reasoning chain:

$$\mathcal{R} = f_{\mathrm{VLM}}(\mathcal{E}, \widetilde{Q}; \theta_1), \tag{18a}$$

$$\mathcal{L}_{\mathrm{R}} = \mathrm{CE}(\mathcal{R}, \mathcal{R}^*), \tag{18b}$$

 doi:10.20944/preprints202605.1648.v1

where $\mathcal{R}$ denotes the generated reasoning chain, $\mathcal{R}^*$ is the ground-truth reasoning chain, and θ_1 denotes the trainable parameters in Stage 1. Prompt Template 3 is used to guide the model to generate step-by-step reasoning grounded in both visual and knowledge evidence.

Prompt Template 3 — Chain-of-Evidence Reasoning

/* Instruction */
Task: Generate a step-by-step reasoning chain
Input:

- Question: `[question]`
- Visual evidence (#id): `[v1]` ... `[vk]`
- Knowledge evidence: `[k1]` ... `[km]`

Please:

1. Integrate the visual and knowledge evidence into the reasoning process
2. Refer to relevant regions when necessary

Output format:

- Reasoning: `[step-by-step reasoning]`

/* Instruction */

Stage 2: Answer Generation

After Stage 1, we fine-tune a separate model to generate answers based on the reasoning chain:

$$\mathcal{A} = f_{VLM}(\mathcal{R}, Q; \theta_2) \tag{19}$$

where θ_2 denotes the parameters fine-tuned with the following instruction:

Prompt Template 4 — Answer Inference

/* Instruction */
Given: `[question]` and the above `[reasoning]`
Output format:

- Answer: `[answer]`

/* Instruction */

The objective is:

$$\mathcal{L}_{MECR} = \mathcal{L}_{\text{answer}}(\mathcal{A}, \mathcal{A}^*) \tag{20}$$

where $\mathcal{A}^*$ is the ground truth answer.

4. Experiments

4.1. Datasets

We evaluate our method on three widely adopted benchmarks for knowledge-based VQA: **OK-VQA** [10], **A-OKVQA** [11], and **F-VQA** [9].

1) ***OK-VQA.*** OK-VQA is a benchmark for knowledge-based VQA, containing 14k image-question pairs drawn from diverse domains, including science, sports, history, and technology. The questions are open-ended, and each one is annotated with ten human-provided answers. The dataset is divided into 9k training questions and 5k testing questions.

2) ***A-OKVQA.*** A-OKVQA extends OK-VQA and includes 25k image-question pairs, requiring models to leverage substantial commonsense and world knowledge for answer prediction. Built on COCO 2017, it consists of 17k training questions, 1.1k validation questions, and 6.7k testing questions. Each question is associated with ten open-ended answers for direct answer (DA) evaluation and four candidate choices for multiple-choice (MC) evaluation. Because the test answers are not publicly released, performance on the test split is measured through leaderboard submission.

3) ***F-VQA.*** F-VQA is a VQA benchmark in which answering the questions requires external knowledge, and each sample is accompanied by supporting fact triplets in addition to the image-question-

 doi:10.20944/preprints202605.1648.v1

answer triplets. In our setting, each textual knowledge triplet is further linked to its corresponding image.

Knowledge Source.

We built a task-specific knowledge base from WIT [47] and ConceptNet [24], using **ConceptNet 5.7** (English dump, October 2020) and the **WIT dataset**[1] (May 2022 release). For ConceptNet, only English triplets (i.e., `/c/en/...`) were retained. For WIT, we extracted the `context_page_description` and `context_section_description` fields, and removed overlong or non-English entries. To ensure task relevance, we kept only items containing at least one token from either (i) the 10,446 visual entities detected by our visual entity linking module on the OK-VQA/A-OKVQA training set, or (ii) the 7,912 non-numeric answers in the training annotations. After lowercasing, punctuation removal, and deduplication, the final knowledge base contains **70,669** entries, including **53,278** from ConceptNet and **17,391** from WIT.

4.2. Performance Metrics

We use ***Acc@1*** and ***Exact Match (EM)*** for evaluation. Given a predicted answer $\hat{y}$, the Acc@1 score is defined as:

$$\text{Acc@1} = \min\left(1, \frac{\#S(y)}{3}\right) \tag{21}$$

For EM, all annotated answers are treated equally:

$$EM(y, S) = \min(\#S(y), 1) \tag{22}$$

A knowledge item is regarded as pseudo-relevant if it contains any human-annotated answer. **PRRecall@K** indicates whether the top-K retrieved knowledge items include at least one pseudo-relevant item:

$$\text{PRRecall@K} = \min\left(\sum_{k=1}^{K} H(kg, S), 1\right) \tag{23}$$

where $H(kg, S) = 1$ when the retrieved knowledge item kg contains any answer in S, and 0 otherwise.

4.3. Implementation Details

Experimental Setup.

All experiments are implemented in PyTorch 2.5.1, using MiniCPM-V 2.6 [42] as the backbone on NVIDIA A40 GPUs. We employ mixed-precision training with gradient checkpointing. Visual regions are detected using RAM [43]. Both semantic and spatial contrastive learning are adopted, and region scores are computed according to Eq. (9). The model is optimized with AdamW using a batch size of 2 and a learning rate of 1.0×10^{-6}. MiniCPM-V is fine-tuned in two stages for reasoning and answer prediction. The maximum input and output lengths are set to 512 and 128 tokens, respectively. Table 1 reports the average runtime, FLOPs, and memory usage of different components.

We benchmark VAMER against 25 VQA methods on the OK-VQA dataset. These baselines are divided into four groups: classification-based methods, methods using explicit knowledge, methods using implicit knowledge, and methods combining both explicit and implicit knowledge. The details are given below.

1) ***Classification-based methods.*** These methods predict answers directly from images and questions without introducing external knowledge. MLP [48] encodes image and question features separately with a fully connected layer and then combines them through a skip-thought GRU for answer prediction. BAN [2] adopts a co-attention mechanism over question features and bottom-up object features extracted from the image. MUTAN [3] models bilinear interactions between visual and textual features through a Tucker-decomposition-based multimodal fusion module. 2) ***Methods leveraging explicit knowledge.*** These approaches retrieve relevant knowledge from external sources to

[1] WIT: https://huggingface.co/datasets/google/wit

 doi:10.20944/preprints202605.1648.v1

enhance reasoning. ConceptBert [12], Learning Knowledge-Tree Retrieval [16] use ConceptNet [24] to retrieve fact-based knowledge and construct a fact graph. Other models, such as REVEAL [49], KRISP [20], VLC-BERT [50], and MAVEx [14], draw knowledge from various sources, including ConceptNet [24], Wikipedia [23], DBpedia [58].

Table 1. Efficiency breakdown of our multi-stage pipeline on MiniCPM-V 2.6. Memory denotes peak usage (shared backbone); time is cumulative.

Method / Stage	Mem. (GB)	FLOPs (G)	Time (s)	OK-VQA (%)
MiniCPM-V (Base)	17.2	28.0	2.15	50.40
VEL	17.7	28.2	1.20	-
KR	0.4	0.6	0.20	-
MECR	17.9	30.8	2.65	-
Ours (Total)	**17.9**	**30.8**	**4.05**	**63.97**

3) ***Methods leveraging implicit knowledge.*** These approaches leverage the implicit knowledge encoded in large pre-trained models, which do not explicitly include external knowledge bases but have been trained on vast datasets. PICA-Full [18] and CBMT5-11B [13] convert images to text using image captioning models and input the generated text into large language models (LLMs) for augmented knowledge. Flamingo [19], OTM [53], PCPA [54] tune a visual encoder for frozen LLMs to enhance multimodal reasoning. MM-COT [22] learns to generate rationales based on ground-truth annotations before producing the final answer by utilizing all available information.

4) ***Methods combining explicit and implicit knowledge.*** Hybrid approaches integrate both explicit and implicit knowledge to enhance reasoning. KAT [27] extends PICa by using Wikipedia [23] as an additional knowledge source and performing ensemble and end-to-end fine-tuning across components. REVIVE [21] incorporates extra object-centric visual features, resulting in improvements over KAT. MMReasoner [28] combines information extracted by multiple vision APIs (e.g., dense captioner, object detector, OCR) with large language models for reasoning and vision-language models (VLMs) for final answer generation.

4.4. Compared with the Baselines

1) ***On OKVQA dataset:*** In Table 2, we present the comprehensive performance evaluation of various methods on the OK-VQA dataset, revealing several insightful trends. Firstly, VAMER achieves the highest performance in both Acc@1 and EM metrics, demonstrating its effectiveness in tackling Knowledge-based VQA challenges. Secondly, compared to classification-based methods such as MLP, BAN, and MUTAN, VAMER exhibits a substantial improvement on Acc@1 and EM scores, with an increase of approximately 43.30% on Acc@1. This indicates that relying solely on attention mechanisms for questions and images is insufficient for solving KB-VQA tasks, emphasizing the critical role of external knowledge. Thirdly, VAMER outperforms the models based on implicit knowledge, such as CBMT5 and PICA, by effectively integrating textual knowledge and visual evidence through multimodal queries with explicit entities, resulting in relatively reliable reasoning. Lastly, VAMER outperforms hybrid methods like KAT, REVIVE, and MM-Reasoner by leveraging the VEL module to extract key visual entities with reduced noise, significantly enhancing the reasoning process. Furthermore, experiments with Qwen2.5VL and LLaVA-1.5-onevision verify the universality of our framework, demonstrating consistent gains on OK-VQA. 2) ***On A-OKVQA dataset:*** A-OKVQA is a large KB-VQA dataset. Table 3 provides a comparative analysis of various models on A-OKVQA, focusing on direct-answer (DA) and multiple-choice (MC) evaluations. The insights drawn from the table are twofold: 1) Compared to traditional pre-trained visual language (VL) models, the proposed VAMER significantly outperforms VL models, such as ViL-BERT, ClipCap and LXMERT. For instance, compared to LXMERT, VAMER significantly improves the DA score (62.57% vs. 41.60%) and the MC score (73.74% vs. 25.90%), highlighting VAMER's advantages in accurate entity anchoring and reliable multimodal evidence reasoning. 2) Compared to hybrid approaches such as KRISP and VLC-BERT,

 doi:10.20944/preprints202605.1648.v1

which incorporate both implicit and explicit knowledge to compensate for the limitations of large models, VAMER adopts a more focused strategy. Hybrid methods typically struggle to identify image entities directly relevant to the question, resulting in distractions from irrelevant knowledge during answer generation. In contrast, VAMER effectively pinpoints the key visual entities mentioned in the question, thereby reducing the influence of extraneous information and improving reasoning accuracy.

Table 2. Performance on the OK-VQA Dataset.

Method	Backbone Size	Knowledge Source	Acc@1	EM
Classification-based methods				
MLP [48]	-	-	20.67	-
BAN [2]	-	-	25.10	-
MUTAN [3]	-	-	26.41	-
Methods leveraging explicit knowledge				
ConceptBert [12]	110M	ConceptNet	33.70	-
LSKR [16]	13B	ConceptNet + Vicuna	56.07	-
KRISP [20]	-	Wikipedia + ConceptNet + VQA P. T	38.90	-
REVEAL [49]	-	WIT + CC12M + Wikipedia + VQA-2	55.22	-
VLC-BERT [50]	-	VQA.P. T + COMET	43.14	-
MAVEX [14]	-	Wikipedia + ConceptNet + Google Image	40.28	41.37
RR-VEL [15]	770M	ConceptNet + Ascent + hasPart	49.48	55.76
CEIK [51]	-	Google Search	55.70	60.50
TRiG-Ensemble [17]	-	Wikipedia	50.50	54.73
RA-VQA-FrDpr [52]	-	Google Search	51.22	55.77
Methods leveraging implicit knowledge				
PICA-Full [18]	175B	Frozen GPT-3	48.00	-
CBMT5-11B [13]	11B	CBMT5	47.90	-
OTM [53]	7B	LLaVA-1.5	61.30	-
PCPA [54]	7B	LLaMA-7b	53.60	-
MiniCPM-V2.6 [42]	8B	MiniCPM-V2.6	51.40	-
LLaVA-1.5-onevision-8B [55]	8B	LLaVA-1.5-onevision-8B	57.06	-
Qwen2.5VL-7B [56]	7B	Qwen2.5VL-7B	55.06	-
Flamingo (80B) [19] (32-shot)	80B	Chinchilla (70B)	57.80	-
Methods combining explicit and implicit knowledge				
KAT [27]	175B	Wikipedia + Frozen GPT-3	53.10	54.40
prophet [57]	175B	MCAN + Frozen GPT-3	61.10	-
REVIVE [21]	175B	Wikipedia + Frozen GPT-3	56.60	58.00
MM-Reasoner [28]	-	Multiple Vision APIs + GPT-4-32k	59.20	60.80
VAMER (MiniCPM-V2.6) (Ours)	8B	**WIT + ConceptNet**	**63.97**	**68.70**
-(LLaVA-1.5-onevision-8B)	8B	**WIT + ConceptNet**	**62.47**	**65.70**
-(Qwen2.5VL-7B)	7B	**WIT + ConceptNet**	**59.65**	**63.54**

 doi:10.20944/preprints202605.1648.v1

Table 3. Performance on the A-OKVQA Dataset (Evaluated on the Test Set). DA Denotes the Direct Answer Evaluation, and MC Denotes the Multiple-Choice Evaluation.

Method	DA	MC
ViL-BERT [59]	41.50	21.90
LXMERT [60]	41.60	25.90
ClipCap [26]	43.80	15.80
KRISP [20]	42.20	27.10
VLC-BERT [50]	38.05	–
CEIK [51]	45.60	–
PCPA [54]	48.50	–
MM-Reasoner [28]	60.20	–
MM-COT [22]	50.57	–
REVEAL [49]	51.50	–
VAMER (Ours)	**62.57**	**73.74**

4.5. Baselines

3) ***On F-VQA dataset:*** In Table 4, we present a comprehensive comparison of various approaches on the F-VQA dataset. VAMER achieves 86.31% on accuracy, outperforming the reported human benchmark under this evaluation protocol (77.99%) and traditional knowledge-based approaches (56.91%-61.10%). To evaluate the robustness of our Visual Entity Linking (VEL) module, we replaced the RAM [43] with mainstream object detectors Faster R-CNN [25] and DETR [62]. The accuracy is slightly lower than RAM-based localization, which still substantially outperforms existing knowledge-driven methods and human performance.

Table 4. Performance on the F-VQA Dataset.

Method	Knowledge Source	Acc@1
Human	-	77.99
F-VQA	ConceptNet	56.91
ZS-VQA [61]	ConceptNet	58.27
F-VQA (Ensemble)	ConceptNet	58.76
MM-Reasoner (Ensemble)	Google Search	61.10
MiniCPM-V	MiniCPM-V	70.11
VAMER (w/RAM)	**MiniCPM-V + ConceptNet**	**86.31**
w/Faster R-CNN [25]	MiniCPM-V + ConceptNet	82.47
w/DETR [62]	MiniCPM-V + ConceptNet	83.65

4.6. Ablation Studies

To evaluate the contribution of each component, we perform ablation studies on OK-VQA, A-OKVQA-Val, and F-VQA. As reported in Table 5, VAMER denotes the complete model. "w/o VEL" removes the visual entity linking module and directly uses the original question for knowledge retrieval. We also study the role of contrastive learning by separately removing semantic contrastive learning ("w/o Semantic CL") and spatial contrastive learning ("w/o Spatial CL"). "w/o MKR" removes the multimodal retriever. "w/o Visual Score" indicates that the visual similarity score is not used, whereas "w/o Text Score" removes the text-based similarity score during knowledge retrieval. Detailed observations are given below. (1) Comparing variants "1-2" shows the importance of the Visual Entity Linking (VEL) module. Without VEL, VAMER drops by 8.46% in Acc@1 and 9.43% in EM on OK-VQA. This result highlights the role of visual anchoring in providing informative

 doi:10.20944/preprints202605.1648.v1

evidence for answer prediction. (2) Based on an analysis of the outcomes from variations "3-4", we assess the effect of Semantic-Spatial contrastive learning in the VEL module. On the OK-VQA dataset, removing semantic contrastive learning and spatial contrastive learning leads to decreases of 5.12% and 4.51% in Acc@1, respectively. The results indicate that both components play important roles in enhancing visual-textual understanding. (3) By examining the ablation of MKR variations "5-7", we have several observations: 1) The rich implicit knowledge within large language models mitigates the impact of removing the Multimodal Knowledge Retriever (MKR), resulting in a slight performance drop (2.81% in Acc@1 on OK-VQA). 2) Removing scoring mechanisms during retrieval causes substantial degradation (11.17% and 13.74% drops in Acc@1 for Visual Score and Text Score, respectively), indicating that retrieving irrelevant knowledge interferes with the model's reasoning process. (4) Comparing the performance of variation "8" reveals that two-stage reasoning outperforms direct answer prediction, with improvements of 7.90% on Acc@1 and 6.53% on EM on OK-VQA. These results demonstrate the benefits of decomposing the reasoning process into multiple steps.

Table 5. Ablation of Key Components in VAMER.

Method	OK-VQA		A-OKVQA (val)		F-VQA
	Acc@1	EM	DA	MC	Acc@1
1. VAMER	**64.07±.85**	**67.60±.73**	**67.05±.95**	**81.45±.89**	**85.81±.88**
Ablation of VEL					
2. w/o VEL	55.61±.91	58.17±.73	61.76±.71	75.99±.88	76.73±.83
3. w/o Semantic CL	58.95±.53	61.33±.85	63.38±.71	77.45±.74	79.01±.95
4. w/o Spatial CL	59.56±.90	61.95±.73	64.03±.82	78.01±.91	80.13±.85
Ablation of MKR					
5. w/o MKR	61.26±.16	65.31±.68	59.93±.61	73.53±.74	70.82±.85
6. w/o Visual Score	52.90±.60	57.56±.93	58.52±.68	72.29±.77	73.24±.51
7. w/o Text Score	50.33±.13	56.15 ±.75	56.15±.77	69.72±.93	71.03±.94
Ablation of MECR					
8. w/o Two-stage	56.17±.82	61.07±.93	62.41±.86	76.33±.94	75.80±.71

4.7. Effectiveness of the Visual Entity Linking (VEL) Module

To validate the effectiveness of the semantic-spatial contrastive learning mechanism in linking key visual entity regions for knowledge retrieval, we conduct ablation studies examining different components of the VEL module across three datasets. As shown in Table 6 (R@ denotes PRRecall@), removing both semantic and spatial contrastive learning in the VEL module ("w/o Semantic & Spatial") results in a significant performance decline. R@1 drops by 8.89%, 6.24%, and 7.15% on OK-VQA, A-OKVQA, and F-VQA, respectively. This demonstrates that without contrastive learning, the VEL module cannot effectively anchor question-relevant visual regions, leading to less accurate knowledge retrieval. Further analysis reveals that semantic contrastive learning plays a crucial role in region localization and knowledge retrieval. Removing it ("w/o Semantic CL") severely hinders the VEL's ability to anchor question-relevant regions, leading to R@1 drops of 9.76%, 5.22%, and 10.78% on the three datasets. In contrast, spatial contrastive learning ("w/o Spatial CL") has a smaller impact on the VEL module, with R@1 dropping by 2.34%, 0.79%, and 7.62%, emphasizing the critical role of semantic contrastive learning in anchoring question-relevant regions. Intuitively, the reason is that the question referents and key visual entities are varied, involving diverse concepts such as animals, sports, food, and transport vehicles. Accurately identifying these diverse entities from both questions and images is challenging, and semantic contrastive learning plays a vital role in training process.

 doi:10.20944/preprints202605.1648.v1

Table 6. Evaluation on different components of the VEL module for retrieval performance.

Method	OK-VQA			A-OKVQA			F-VQA		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
w/o Sem. & Spa.	58.32	71.45	76.89	59.87	71.93	78.45	69.41	75.68	80.12
w/o Sem. CL	57.45	69.21	77.34	60.89	70.45	77.91	65.78	78.92	82.45
w/o Spa. CL	64.87	79.45	83.21	65.32	72.12	84.56	68.94	82.45	85.67
VEL	67.21	83.74	87.13	66.11	73.96	84.77	76.56	88.89	90.02

4.8. Impact of Scene Complexity

We conduct several experiments to analyze how scene complexity affects the performance of the proposed method, focusing on the impact of visual information redundancy as reflected by the number of object regions in the scene. Specifically, images containing more object regions typically exhibit higher redundancy, with additional irrelevant visual regions beyond the key ones. To evaluate the impact of redundancy, we divided the test dataset into three subsets based on the number of object regions in each image: 1-3 object regions (simple scenes), 4-9 object regions, and 9+ object regions. The experimental results in Figure 4 demonstrate that the VEL module consistently enhances retrieval performance across all object number ranges and datasets. For images with 1-3 object regions, the VEL module achieves the highest improvements on R@5, with gains of 7.29%, 7.32%, and 8.80% on OK-VQA, A-OKVQA, and F-VQA, respectively. The performance improvements remain strong for scenes with 4-9 object regions, showing gains of 7.33%, 7.78%, and 8.44% on R@5. Even in complex scenes with more than 9 object regions, the VEL module maintains substantial improvements (6.89%, 7.34%, and 8.32% for R@5), highlighting its effectiveness across a wide range of scene complexities. In addition to scene complexity analysis, another critical aspect lies in the knowledge retrieval component. The Multimodal Knowledge Retriever (MKR) module in VAMER aims to obtain knowledge candidates as support for visual reasoning.

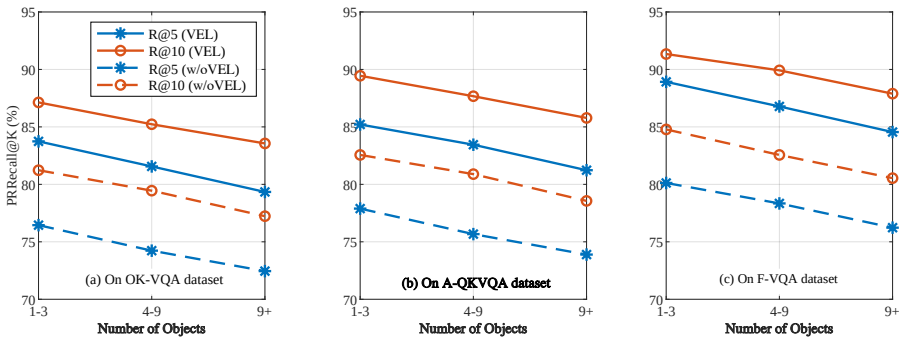


Figure 4. Effect of VEL-generated queries.

4.9. Knowledge Utility and Robustness Analysis

To assess the quality of retrieved knowledge and the robustness of the system against retrieval errors, we conducted a controlled study on 300 random samples from the A-OKVQA validation set, following the protocol of [51].

Protocol

Each question is evaluated under two settings: (i) an *implicit path*, where the model predicts y_{k+1} without using external knowledge; and (ii) *explicit paths*, where each y_j is predicted based on a single retrieved item s_j. By comparing y_j with y_{k+1} and the ground-truth answer set A, each retrieved item is categorized as follows:

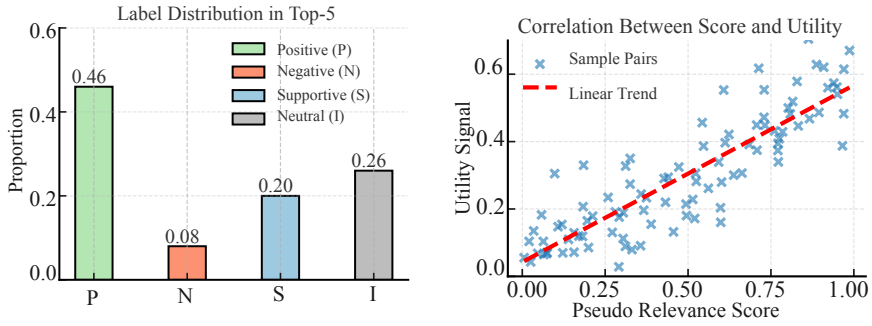
- **Positive (P)**: $y_{k+1} \notin A, y_j \in A$, and $H(s_j, A) = 1$
- **Negative (N)**: $y_{k+1} \in A, y_j \notin A$

 doi:10.20944/preprints202605.1648.v1

- **Supportive (S)**: $y_{k+1} \in A$, $y_j \in A$, and $H(s_j, A) = 1$
- **Neutral**: All other cases

Metrics

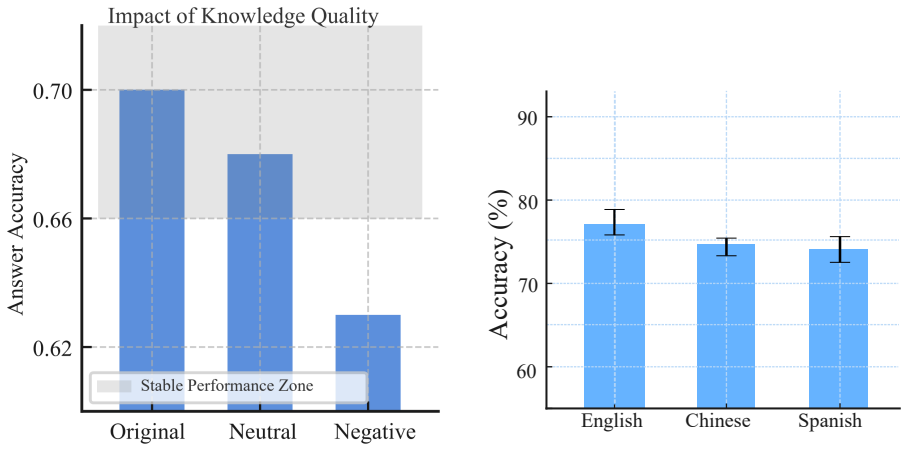
We report two perspectives: *Utility* considers only **P** cases as positive, measuring the ability of retrieved knowledge to correct the model; *Relevance* considers **P** and **S** as positives, evaluating alignment with the ground-truth regardless of whether the final prediction changes. As shown in Figure 5(a), the majority of top-ranked retrieved knowledge instances are **Positive** or **Supportive**, indicating that the retrieved evidence is either corrective or aligned with the ground truth. As shown in Figure 5(b), the positive correlation between pseudo relevance and utility demonstrates that our scoring strategy effectively prioritizes useful knowledge.



(a) Label distribution. (b) Score-Utility correlation.

Figure 5. Analysis of retrieved knowledge utility.

To further evaluate the robustness of our system under noisy or mismatched retrieval, we simulate adverse conditions by replacing the top-1 retrieved knowledge with either a **Neutral** or a **Negative** instance. As shown in Figure 6(a), these substitutions result in accuracy drops of 2% and 7%, respectively. This highlights the sensitivity of answer prediction to irrelevant or misleading knowledge. Accuracy stays stable under neutral retrieval, suggesting that relevance-aware retrieval and visual grounding mitigate errors arising from imperfect knowledge. In addition, we evaluate **cross-lingual generalization** by testing VAMER on the A-OKVQA-VAL dataset in English, Chinese, and Spanish. As shown in Figure 6(b), the model maintains strong accuracy across languages—82.7% in English, 79.0% in Chinese, and 78.5% in Spanish—demonstrating its versatility across linguistic contexts.



(a) Noisy knowledge robustness. (b) Generalization across languages.

Figure 6. Robustness-Multilingual.

 doi:10.20944/preprints202605.1648.v1

We evaluate rationale quality using two complementary metrics: ROUGE-L for lexical overlap and SBERT cosine similarity for semantic alignment, following the protocol in [22,63]. As shown in Table 7, our MECR module consistently outperforms the baseline on both measures. It achieves a ROUGE-L score of 35.29% and an SBERT score of 68.34%, compared to 30.57% and 54.83% from MiniCPM-V. These results demonstrate that our method produces explanations that are both more textually aligned and semantically closer to human reasoning.

Table 7. Reasoning quality evaluation on A-OKVQA.

Method	ROUGE-L↑(%)	SBERT↑(%)
MiniCPM-V	30.57	54.83
Ours (MECR)	**35.29**	**68.34**

4.10. Case Study

We present the qualitative examples in Figure 7, and provide the analysis for failure case study. The examples in the first two rows of Figure 7 (top) show successful cases of VAMER, which accurately links key entities in questions with corresponding detected object regions and effectively manages their interrelationships. For instance, in the first example, VAMER correctly identifies an orange container in a baseball scene and retrieves relevant knowledge based on the composed regions.

Although the design of our visual entity linking module mitigates some semantic ambiguities, the model still struggles to capture deeper intent. In the failure case, the proposed model predicts "Uniqlo" instead of the correct answer "Polo", the match sponsor. This misprediction arises from the model's reliance on surface-level visual cues such as clothing style and color patterns, which are commonly associated with "Uniqlo" in the training data. However, the distinctive brand logos or textual hints indicating "Polo" are either absent or visually inconspicuous. This case reveals a limitation of the model in visually ambiguous scenes, where it overfits to familiar patterns rather than identifying fine-grained brand cues.

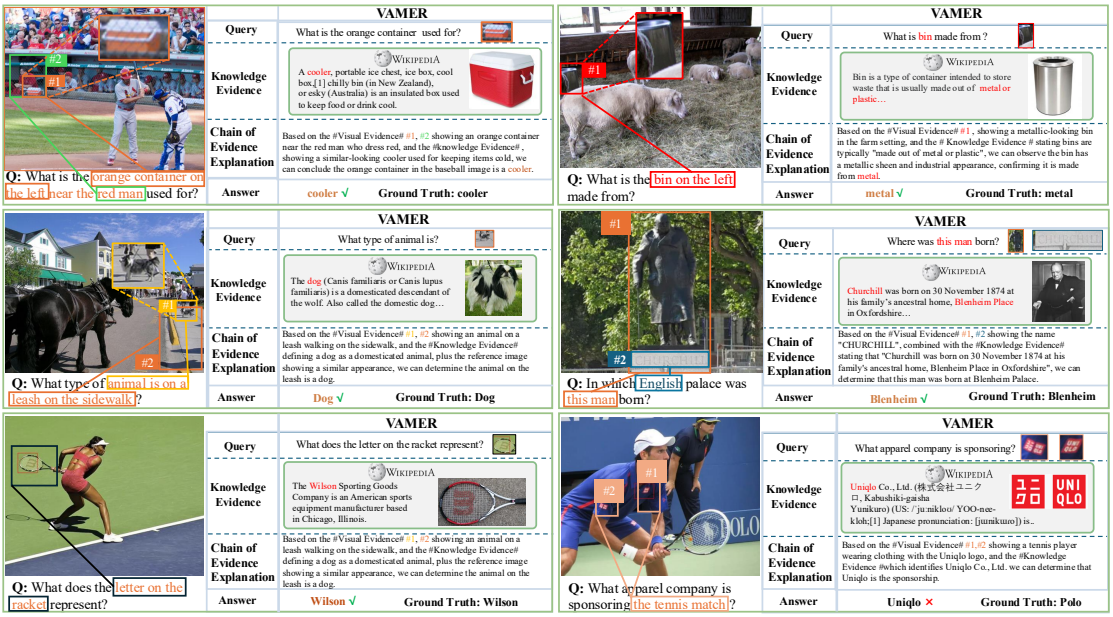


Figure 7. Success and failure cases of VAMER on the OK-VQA dataset.

5. Conclusions

In this paper, we propose VAMER, a novel framework to address the visual entity anchoring and visual-aware reasoning challenges in knowledge-based VQA. To tackle the visual entity anchoring issue, we design a Visual Entity Linking (VEL) module that leverages semantic and spatial information

 doi:10.20944/preprints202605.1648.v1

to locate relevant regions in the image, enabling precise multimodal query generation. To address the visual-aware reasoning issue, we introduce a multimodal evidence chain reasoning (MECR) module that constructs reasoning paths by interleaving visual and textual information, improving the reliability of inference chains. Our framework achieves outstanding performance on OK-VQA, A-OKVQA, and F-VQA datasets. Ablation studies further validate the effectiveness of each proposed component.

Acknowledgments: This work was supported by Grant Number JPMJSP2132.

References

1. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," in *ICCV*, 2015, pp. 2425–2433.
2. J. Kim, J. Jun, and B. Zhang, "Bilinear attention networks," in *NeurIPS*, 2018, pp. 1571–1581.
3. H. Ben-Younes, R. Cadène, M. Cord, and N. Thome, "MUTAN: multimodal tucker fusion for visual question answering," in *ICCV*, 2017, pp. 2631–2639.
4. B. Qin, H. Hu, and Y. Zhuang, "Deep residual weight-sharing attention network with low-rank attention for visual question answering," *IEEE Transactions on Multimedia*, vol. 25, pp. 4282–4295, 2023.
5. T. Qian, J. Chen, S. Chen, B. Wu, and Y.-G. Jiang, "Scene graph refinement network for visual question answering," *IEEE Transactions on Multimedia*, pp. 3950–3961, 2023.
6. H. Zhong, J. Chen, C. Shen, H. Zhang, J. Huang, and X.-S. Hua, "Self-adaptive neural module transformer for visual question answering," *IEEE Transactions on Multimedia*, vol. 23, pp. 1264–1273, 2021.
7. F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Visual question answering with dense inter- and intra-modality interactions," *IEEE Transactions on Multimedia*, vol. 23, pp. 3518–3529, 2021.
8. Z.-X. Jin, H. Wu, C. Yang, F. Zhou, J. Qin, L. Xiao, and X.-C. Yin, "Ruart: A novel text-centered solution for text-based visual question answering," *IEEE Transactions on Multimedia*, pp. 1–12, 2023.
9. P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "Fvqa: Fact-based visual question answering," *TPAMI*, pp. 2413–2427, 2018.
10. K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A visual question answering benchmark requiring external knowledge," in *CVPR*, 2019, pp. 3195–3204.
11. D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-OKVQA: a benchmark for visual question answering using world knowledge," in *ECCV*, 2022, pp. 146–162.
12. F. Gardères, M. Ziaeefard, B. Abeloos, and F. Lécué, "Conceptbert: Concept-aware representation for visual question answering," in *EMNLP*, 2020, pp. 489–498.
13. A. Salaberria, G. Azkune, O. L. de Lacalle, A. Soroa, and E. Agirre, "Image captioning for effective use of language models in knowledge-based visual question answering," *Expert Syst. Appl.*, p. 118669, 2023.
14. J. Wu, J. Lu, A. Sabharwal, and R. Mottaghi, "Multi-modal answer validation for knowledge-based VQA," in *AAAI*. AAAI Press, 2022, pp. 2712–2721.
15. J. You, Z. Yang, Q. Li, and W. Liu, "A retriever-reader framework with visual entity linking for knowledge-based visual question answering," in *ICME*, 2023, pp. 13–18.
16. N. Xu, Z. Lu, H. Tian, R. Kang, J. Cao, Y. Zhang, and A. Liu, "Learning to supervise knowledge retrieval over a tree structure for visual question answering," *IEEE Transactions on Multimedia*, vol. 26, pp. 6689–6700, 2024.
17. F. Gao, Q. Ping, G. Thattai, A. Reganti, Y. N. Wu, and P. Natarajan, "Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering," in *CVPR*, 2022, pp. 5057–5067.
18. Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, "An empirical study of GPT-3 for few-shot knowledge-based VQA," in *AAAI*, 2022, pp. 3081–3089.
19. J. Alayrac, J. Donahue, P. Luc, and K. S. etc, "Flamingo: a visual language model for few-shot learning," in *NeurIPS*, 2022.
20. K. Marino, X. Chen, D. Parikh, A. Gupta, and M. Rohrbach, "KRISP: integrating implicit and symbolic knowledge for open-domain knowledge-based VQA," in *CVPR*, 2021, pp. 14 111–14 121.
21. Y. Lin, Y. Xie, D. Chen, Y. Xu, C. Zhu, and L. Yuan, "REVIVE: regional visual representation matters in knowledge-based visual question answering," in *NeurIPS*, 2022.
22. Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *Trans. Mach. Learn. Res.*, vol. 2024, 2024.
23. D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, p. 78–85, 2014.

 doi:10.20944/preprints202605.1648.v1

24. R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," *AAAI*, p. 4444–4451, 2017.
25. S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.
26. R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: CLIP prefix for image captioning," *CoRR*, 2021.
27. L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "KAT: A knowledge augmented transformer for vision-and-language," in *NAACL*, 2022, pp. 956–968.
28. M. Khademi, Z. Yang, F. Frujeri, and C. Zhu, "Mm-reasoner: A multi-modal knowledge-aware framework for knowledge-based visual question answering," in *EMNLP*, 2023, pp. 6571–6581.
29. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.
30. T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3375–3385, 2023.
31. A. N. Venkitasubramanian, T. Tuytelaars, and M. Moens, "Entity linking across vision and language," *Multim. Tools Appl.*, pp. 22 599–22 622, 2017.
32. S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity disambiguation for noisy social media posts," in *ACL*, 2018, pp. 2000–2008.
33. Q. Zheng, H. Wen, M. Wang, and G. Qi, "Visual entity linking via multi-modal learning," *Data Intell.*, pp. 1–19, 2022.
34. J. Gan, J. Luo, N. Wang, S. Wang, W. He, and Q. Huang, "Multimodal entity linking: A new dataset and A baseline," in *MM*, 2021, pp. 993–1001.
35. J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *CVPR*, 2017, pp. 1988–1997.
36. M. F. Ishmam, M. S. H. Shovon, M. F. Mridha, and N. Dey, "From image to language: A critical analysis of visual question answering (VQA) approaches, challenges, and opportunities," *Inf. Fusion*, vol. 106, p. 102270, 2024.
37. H. Song, L. Dong, W. Zhang, T. Liu, and F. Wei, "CLIP models are few-shot learners: Empirical studies on VQA and visual entailment," in *ACL*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., 2022, pp. 6088–6100.
38. S. Amizadeh, H. Palangi, A. Polozov, Y. Huang, and K. Koishida, "Neuro-symbolic visual reasoning: Disentangling "visual" from "reasoning"," in *ICML*, vol. 119, 2020, pp. 279–290.
39. M. Choi, H. Goel, M. Omama, Y. Yang, S. Shah, and S. Chinchali, "Towards neuro-symbolic video understanding," in *ECCV*, vol. 15136, 2024, pp. 220–236.
40. D. Chen, J. Liu, W. Dai, and B. Wang, "Visual instruction tuning with polite flamingo," in *AAAI*, 2024, pp. 17 745–17 753.
41. T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," in *CVPR*, 2023, pp. 14 953–14 962.
42. Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, and H. Zhu, "Minicpm-v: A GPT-4V level MLLM on your phone," *CoRR*, vol. abs/2408.01800, 2024.
43. Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu, Y. Guo, and L. Zhang, "Recognize anything: A strong image tagging model," *CoRR*, 2023.
44. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.
45. J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," L. Ku, A. Martins, and V. Srikumar, Eds. ACL, 2024, pp. 2318–2335.
46. V. Karpukhin, B. Oguz, S. Min, P. S. H. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in *EMNLP*, 2020, pp. 6769–6781.
47. K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," *CoRR*, 2021.
48. Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nat.*, pp. 436–444, 2015.
49. Z. Hu, A. Iscen, C. Sun, Z. Wang, K. Chang, Y. Sun, C. Schmid, D. A. Ross, and A. Fathi, "Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory," in *CVPR*, 2023, pp. 23 369–23 379.

 doi:10.20944/preprints202605.1648.v1

50. S. Ravi, A. Chinchure, L. Sigal, R. Liao, and V. Shwartz, "VLC-BERT: visual question answering with contextualized commonsense knowledge," in *WACV*, 2023, pp. 1155–1165.
51. Q. Wang, J. Liu, and W. Wu, "Coordinating explicit and implicit knowledge for knowledge-based VQA," *Pattern Recognit.*, vol. 151, p. 110368, 2024.
52. W. Lin and B. Byrne, "Retrieval augmented visual question answering with outside knowledge," in *EMNLP*, 2022, pp. 11 238–11 254.
53. H. Ji, Q. Si, Z. Lin, Y. Cao, and W. Wang, "Towards one-to-many visual question answering," in *EMNLP*, 2024, pp. 16 931–16 943.
54. Z. Hu, P. Yang, Y. Jiang, and Z. Bai, "Prompting large language model with context and pre-answer for knowledge-based VQA," *Pattern Recognit.*, vol. 151, p. 110399, 2024.
55. X. An, Y. Xie, K. Yang, W. Zhang, X. Zhao, Z. Cheng, Y. Wang, S. Xu, C. Chen, D. Zhu *et al.*, "Llava-onevision-1.5: Fully open framework for democratized multimodal training," *arXiv preprint arXiv:2509.23661*, 2025.
56. A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, and C. Z. and, "Qwen2 technical report," *CoRR*, vol. abs/2407.10759, 2024.
57. Z. Shao, Z. Yu, M. Wang, and J. Yu, "Prompting large language models with answer heuristics for knowledge-based visual question answering," in *CVPR*. IEEE, 2023, pp. 14 974–14 983.
58. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "Dbpedia: A nucleus for a web of open data," in *ISWC*, 2007, pp. 722–735.
59. J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019, pp. 13–23.
60. H. Tan and M. Bansal, "LXMERT: learning cross-modality encoder representations from transformers," in *IJCNLP*, 2019, pp. 5099–5110.
61. Z. Chen, J. Chen, Y. Geng, J. Z. Pan, Z. Yuan, and H. Chen, "Zero-shot visual question answering using knowledge graph," in *ISWC*, ser. Lecture Notes in Computer Science, vol. 12922. Springer, 2021, pp. 146–162.
62. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, vol. 12346, 2020, pp. 213–229.
63. Z. Yang, J. Lin, Z. Guo, Y. Li, X. Li, Q. Li, and W. Liu, "Towards rumor detection with multi-granularity evidences: A dataset and benchmark," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 7188–7200, 2024.