# VAR: MINING VISUAL ANCHORS FROM VISION ENCODERS FOR ENTITY RETRIEVAL IN KNOWLEDGE-BASED VISUAL QUESTION ANSWERING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

In Knowledge-Based Visual Question Answering (KB-VQA), correct answer generation relies on the evidence contained in the retrieved gold document. Therefore, accurately retrieving the gold document is critical. This requires matching the question to entities in the knowledge base. However, standard vision encoders often underemphasize subtle, entity-discriminative cues, entangling them with irrelevant background noise in the final global representation. While recent methods address this via query-side region cropping or knowledge-base-side captioning, they either rely on brittle explicit region selection or incur substantial preprocessing overhead. Moreover, existing verification strategies often rely on binary relevance decisions, discarding the relative strength of visual support among candidates. In this paper, we reveal an overlooked phenomenon: fine-grained, entity-discriminative semantics already persist in high-level token representations, merely underexpressed during global aggregation. We conceptualize these compact, retrieval-critical token subsets as **Visual Anchors**. Based on this insight, we propose **VAR** (**V**isual **A**nchors for Entity **R**etrieval), a fine-grained entity retrieval framework. VAR first mines visual anchors and uses them to guide `[CLS]` updates through the remaining frozen encoder layers, strengthening their contribution to the retrieval representation without pixel-space region extraction or knowledge-base index reconstruction. To further improve gold-document ranking, VAR leverages concise visual-attribute descriptions from Top-ranked candidates. Beyond binary verification, VAR uses a VLM calibrator to estimate continuous visual support and trains a lightweight gate to jointly calibrate the relative visual support of a candidate set against their retrieval confidence for Top-1 document correction. Extensive experiments demonstrate that VAR achieves state-of-the-art performance on E-VQA, InfoSeek, and OK-VQA, showing the effectiveness and efficiency of our method. Code will be released upon acceptance.

## 1 INTRODUCTION

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in Visual Question Answering (VQA) tasks. However, due to the static nature of their training, these models struggle to address highly time-sensitive questions. To address this limitation, Knowledge-Based VQA (KB-VQA) has been proposed to answer fine-grained or specialized questions by retrieving external knowledge (Mensink et al., 2023; Chen et al., 2023). Specifically, in the retrieval phase, existing methods encode both query and knowledge-base images using pretrained vision encoders and retrieve entities based on representation similarity (Yan & Xie, 2024; Lin et al., 2024; Yang et al., 2025a). However, since CLIP-based encoders are primarily trained for vision–language alignment (Radford et al., 2021), their global representations may underemphasize subtle cues required for entity-level discrimination (Tong et al., 2024). During this process, background context becomes entangled with the entities within the learned representations, thus introducing irrelevant noise that compromises retrieval precision.

To address this issue, recent works seek finer-grained visual evidence for retrieval through two main strategies. On the query side, methods (Jian et al., 2024; Hu & Tu, 2026) crop question-relevant regions, construct multi-scale views, or re-encode local patches to suppress the irrelevant context.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

On the knowledge-base side, approaches (Deng et al., 2025; Yang et al., 2025a; Meng et al., 2025) enrich candidate representations with detailed captions or refined visual features. Nevertheless, we contend that fine-grained representations capable of distinguishing entities are already embedded within higher-level feature spaces.

To validate this hypothesis, we conduct an analysis of the representations extracted by CLIP. Our analysis reveals that useful entity-discriminative evidence for retrieval already persists in the encoder yet insufficiently expressed by the default global readout. We conceptualize these compact, highly discriminative, and retrieval-critical token subsets in the high-level embedding space as **Visual Anchors**. Given that Visual Anchors aggregate high-level semantic information that has already been contextualized by the encoder, the reconstructed retrieval representation preserves the inherent global semantic structure.

Query-side enhancement, however, may still leave ambiguity among top-ranked candidates. Passing knowledge from competing entities to the answer generator can introduce conflicting evidence and degrade reasoning (Hong et al., 2026). Existing methods (Zhang et al., 2026; Cocchi et al., 2025) perform evidence verification by learning passage-level binary relevance labels (e.g., `[REL]` / `[NOREL]`) to filter or re-rank retrieved knowledge. Such answer-oriented verification mainly assesses whether individual passages support answering the question, but does not establish whether the source document corresponds to the depicted entity. More importantly, pointwise binary decisions do not directly optimize cross-candidate ranking and discard differences in support strength among candidates. This motivates us to focus comparison and calibration of continuous visual support among candidate set for Top-1 document selection.

Based on the above considerations, we propose **VAR** (**V**isual **A**nchors for Entity **R**etrieval), a fine-grained entity retrieval framework for KB-VQA. VAR consists of two stages: **Visual Anchor Mining (VAM)** and **Description-Grounded Evidence Calibration (DEC)**. **For retrieval**, VAM mines spatially coherent visual anchors from high-level token representations and uses them to guide `[CLS]` updates through the remaining frozen encoder layers, strengthening their contribution to the retrieval representation. **For the retrieval content,** DEC extracts visual-attribute descriptions from a set of high-ranked candidate pages and uses a VLM calibrator to verify whether the candidates correspond to the depicted target entity. The support–reject logit difference is retained as a continuous visual-support signal. A lightweight gate jointly calibrates relative visual support and retrieval confidence with respect to the front-end winner to select the final Top-1 document. Our main contributions are summarized as follows:

- An overlooked phenomenon in KB-VQA retrieval is identified: entity-discriminative information can persist in high-level vision tokens yet be unexpressed in the final global embedding. We conceptualize these compact, highly discriminative, retrieval-relevant token groups as **Visual Anchors**.
- We propose **VAR**, which mines visual anchors with a lightweight selector and uses them to guide `[CLS]` updates through the remaining frozen encoder layers, without pixel-space region extraction or knowledge-base index reconstruction.
- We exploit visual-attribute descriptions to verify whether candidate documents correspond to the depicted target entity. A lightweight gate jointly calibrates continuous visual support and front-end retrieval confidence among a candidate set for robust Top-1 document correction.
- Extensive experiments demonstrate state-of-the-art performance of our method on E-VQA, InfoSeek, and OK-VQA.

## 2 Related Work

### 2.1 Knowledge-Based Visual Question Answering

Knowledge-Based Visual Question Answering (KB-VQA) requires external factual or commonsense knowledge beyond the visual content of the image (Wang et al., 2018; Marino et al., 2019; Ma et al., 2026b;a). Early approaches integrate symbolic and implicit knowledge (Marino et al., 2021), formulate KB-VQA as retrieval followed by answer prediction (Luo et al., 2021), or introduce entity-focused retrieval (Wu & Mooney, 2022). More recent encyclopedic pipelines explore

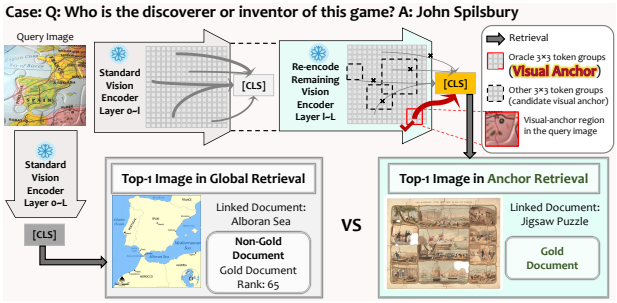



Figure 1: A representative Visual Anchor that promotes the gold document from Rank 65 to Rank 1.

Table 1: **Oracle retrieval performance across token-group scales on InfoSeek.** $\Delta$R@1 is relative to the global representation.

| Token Scale | R@1 | R@5 | R@10 | $\Delta$R@1 |
|---|---|---|---|---|
| 3×3 | **61.4** | **76.6** | **81.9** | **+22.3** |
| 5×5 | 54.2 | 72.8 | 78.7 | +15.1 |
| 7×7 | 49.5 | 70.4 | 76.7 | +10.4 |
| 9×9 | 46.6 | 68.5 | 74.8 | +7.5 |
| 12×12 | 42.3 | 65.3 | 72.7 | +3.2 |
| Global (16×16) | 39.1 | 63.6 | 70.2 | – |

hierarchical retrieval (Caffagni et al., 2024), multimodal reranking (Yan & Xie, 2024), retrieval refinement (Hong et al., 2025), and relevance-aware reflection (Zhang et al., 2026; Cocchi et al., 2025). Despite this progress, fine-grained entity retrieval remains challenging when global visual representations underemphasize subtle entity-discriminative cues.

### 2.2 Fine-Grained Visual Evidence Acquisition and Retrieval Refinement

Fine-grained visual evidence is essential for distinguishing visually similar entities associated with different knowledge sources. REVIVE (Lin et al., 2022) incorporates object-region features, while QACap (Yang et al., 2025b) generates question-aware visual descriptions. Eyes Wide Shut? (Tong et al., 2024) further reveals the limitations of CLIP-style global representations in preserving discriminative visual details. More recent methods explicitly localize or enhance retrieval-relevant visual evidence: Visual Perception Tokens (Yu et al., 2025) performs selective visual perception, Region-R1 (Hu & Tu, 2026) adopts question-conditioned region cropping, and OMGM (Yang et al., 2025a) and KBMR (Xu et al., 2026) exploit multi-granularity or entity-aligned representations. These approaches primarily acquire fine-grained evidence through pixel-space localization or external representation enhancement. In contrast, **VAR** mines **Visual Anchors** from contextualized high-level tokens and uses them to guide `[CLS]` updates through the remaining frozen encoder layers, without pixel-space region extraction or knowledge-base index reconstruction.

## 3 Observation: Visual Anchors in Vision Encoders

In this section, we investigate whether retrieval-discriminative evidence is already encoded in high-level vision tokens but underrepresented in the final global representation. To examine this hypothesis, at a high encoder layer, patch token states from the natural forward pass are replayed through the remaining frozen blocks while updating only the retrieval `[CLS]` token. During replay, its patch-attention mass is concentrated on one spatially coherent token group, allowing us to examine how different internal visual evidence affects the final retrieval representation. As shown in Figure 1, the original global representation is dominated by map-related semantics and ranks the gold document **Jigsaw puzzle** at Rank 65. Conditioning `[CLS]` on a token group covering the puzzle-piece boundaries instead promotes the gold document to Rank 1. This example suggests that entity-discriminative evidence can already persist in high-level tokens, yet receive insufficient emphasis in the default global readout.

To assess the prevalence of this phenomenon, we conduct an Oracle study on 1K randomly sampled InfoSeek examples. Across all token-group scales, we keep the candidate locations fixed and vary only the spatial extent of each group. For each scale, the Oracle selects the candidate that ranks the gold document highest. As shown in Table 1, 3×3 groups improve R@1 from 39.1% to 61.4%, while performance decreases steadily as the group size increases, reaching 42.3% at 12×12. These results indicate that retrieval-discriminative evidence is often concentrated in compact spatial supports. We refer to such compact, highly discriminative, and retrieval-relevant token groups in the high-level embedding space as **Visual Anchors**. These observations motivate VAR to mine and reorganize such internal evidence for fine-grained entity retrieval.

# 4 METHOD

## 4.1 OVERVIEW

In the KB-VQA setting, given a query image $I$ and a question $Q$, the model retrieves external knowledge from a large-scale collection of entity documents $\{D_i\}$ and associated reference images $\{I_i\}$. Motivated by Section 3, we propose Visual Anchors for Entity Retrieval (VAR), consisting of two stages. (1) Visual Anchor Mining (VAM) selects visual anchors from high-level tokens of the frozen vision encoder, constructs anchor-conditioned queries, and aggregates multi-anchor retrieval evidence to rank candidate documents. (2) **Description-Grounded Evidence Calibration (DEC)** uses a VLM calibrator to assess candidate visual descriptions against the query image. A lightweight gate then combines the continuous support–reject margins with front-end retrieval confidence to refine the Top-1 decision. Finally, a lightweight text section reranker selects the most relevant sections from the final Top-1 document for answer generation.

## 4.2 VISUAL ANCHOR MINING

In contrast to methods that extract local regions in pixel space, Visual Anchor Mining (VAM) operates directly on high-level vision tokens. As illustrated in Figure 2, VAM first selects spatially coherent token groups as visual anchors. It then constructs anchor-conditioned query representations for retrieval and aggregates the resulting multi-anchor evidence to score candidate documents.

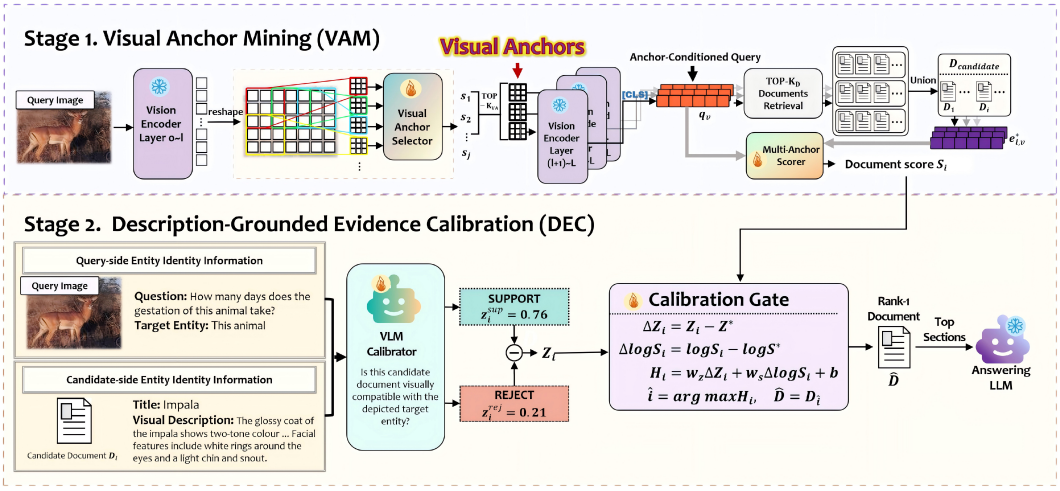



Figure 2: **Overview of VAR.** VAM mines visual anchors and aggregates anchor-conditioned retrieval evidence to rank candidate documents. DEC then combines description-grounded visual support with front-end retrieval confidence to calibrate the Top-1 decision, after which relevant sections are selected for answer generation.

**(1) Visual Anchor Selection.** Given a query image $I$, the first $l$ layers of the frozen vision encoder produce:

$$c, X = f_{\text{body}}(I), \tag{1}$$

where $c \in \mathbb{R}^d$ is the class token and $X = [x_1, \ldots, x_N] \in \mathbb{R}^{N \times d}$ denotes the high-level patch tokens. To capture spatially coherent visual semantics, we define $J$ rectangular regions $\{B_j\}_{j=1}^{J}$ on the token grid, each inducing a spatially coherent token group

$$G_j = \{x_p \in X \mid (u_p, v_p) \in B_j\}, \qquad G_j \subseteq X. \tag{2}$$

where $(u_p, v_p)$ denotes the grid coordinate of token $x_p$. Each $G_j$ therefore forms a candidate visual anchor composed of neighboring high-level tokens.

To estimate the retrieval utility of each candidate visual anchor, we project the high-level patch tokens into a common feature space and extract the tokens within $G_j$:

$$\begin{aligned}
&\bar{X} = W_X(X), \qquad F_j = \left[\bar{X}\right]_{G_j}, \\
&F'_j = \rho_1(\mathcal{C}_{\text{in}}(F_j)), F''_j = F'_j + \rho_2\big(\mathcal{C}_{\text{loc}}(F'_j)\big), \\
&s_j = \text{MLP}\big(\text{AvgPool}\big(\rho_3(\mathcal{C}_{\text{out}}(F''_j))\big)\big),
\end{aligned} \tag{3}$$

where $W_X$ denotes the linear projection. $\mathcal{C}_{\text{in}}$ and $\mathcal{C}_{\text{out}}$ are $1\times1$ convolutions for channel projection, and $\mathcal{C}_{\text{loc}}$ is a $3\times3$ convolution that captures interactions among neighboring tokens. $\rho_1$, $\rho_2$, and $\rho_3$ denote normalization–activation blocks. The resulting score $s_j$ measures the retrieval utility of candidate visual anchor $G_j$. Since a single anchor may capture only one discriminative aspect of an entity, we retain Top-$\text{K}_{\text{VA}}$ anchors:

$$\mathcal{V} = \text{Top-}\,\text{K}_{\text{VA}}\left(\{s_j\}_{j=1}^{J}\right), \tag{4}$$

where $\mathcal{V} \subseteq \{1, \ldots, J\}$ contains the indices of the selected visual anchors.

**(2) Visual-Anchor-Conditioned Retrieval and Multi-Anchor Scoring.** The selected anchors are used to construct complementary query representations, retrieve candidate documents, and aggregate evidence across anchors for document ranking.

• **Anchor-Conditioned Query Construction.** For each selected anchor $G_v$, $v \in \mathcal{V}$, we replay and update only `[CLS]` through the remaining frozen encoder layers while reusing the layer-wise patch token states cached from the natural forward pass. At each replayed layer, let $\alpha_p$ denote the original `[CLS]`-to-patch attention and $\mu = \sum_{p=1}^{N} \alpha_p$ its total patch-attention mass. We concentrate this mass on $G_v$:

$$\widetilde{\alpha}_p^{(v)} = \begin{cases} \mu \dfrac{\alpha_p}{\sum_{r:\, x_r \in G_v} \alpha_r}, & x_p \in G_v, \\ 0, & x_p \notin G_v, \end{cases} \qquad q_v = f_{\text{tail}}(c, X; \widetilde{\alpha}^{(v)}), \tag{5}$$

where $f_{\text{tail}}$ denotes the remaining frozen encoder layers. The normalization preserves the total patch-attention mass $\mu$ while concentrating it on $G_v$, yielding the anchor-conditioned retrieval representation $q_v$.

• **Multi-Anchor Retrieval and Scoring.** Each anchor-conditioned query independently retrieves its Top-$\text{K}_D$ documents from the fixed knowledge-base image index. Their union forms the candidate pool:

$$\mathcal{Q} = \{q_v\}_{v \in \mathcal{V}}, \qquad D_{\text{candidate}} = \bigcup_{q_v \in \mathcal{Q}} \text{Top-}\,\text{K}_D\left(\text{Retrieve}(q_v)\right). \tag{6}$$

For each candidate document $D_i \in D_{\text{candidate}}$, let $e_i$ denote the set of fixed embeddings of its associated knowledge-base images. Under each anchor-conditioned query $q_v$, we select the best-matching image embedding:

$$e^*_{i,v} = \arg\max_{e \in e_i} q_v^\top e. \tag{7}$$

The interaction between the query and the selected image embedding is encoded as:

$$h_{i,v} = \text{LayerNormalization}\big(W_\odot(q_v \odot e^*_{i,v}) + W_\Delta|q_v - e^*_{i,v}| + W_q q_v + W_e e^*_{i,v} + W_x x_{i,v}\big), \tag{8}$$

where $W_\odot$, $W_\Delta$, $W_q$, $W_e$, and $W_x$ denote learnable projection. $x_{i,v}$ summarizes the retrieval information for document $D_i$ under anchor $v$. We then aggregate evidence from all selected anchors to obtain the candidate document score:

$$S_i = \text{softmax}_i(f_{\mathcal{T}}(\{h_{i,v}\}_{q_v \in \mathcal{Q}}, \gamma_i)), \tag{9}$$

where $f_{\mathcal{T}}$ is a lightweight self-attention Transformer and $\gamma_i$ summarizes cross-anchor retrieval information. $\text{softmax}_i$ returns the normalized confidence of candidate $D_i$ over the entire candidate pool. The features $x_{i,v}$ and $\gamma_i$ are detailed in Appendix B.1. Ranked by $S_i$, the Top-$\text{K}_{\text{CA}}$ candidate documents are passed to DEC for subsequent calibration.

### 4.3 DESCRIPTION-GROUNDED EVIDENCE CALIBRATION

To improve gold-document ranking, Description-Grounded Evidence Calibration (DEC) refines the Top-1 decision among the Top-$K_{CA}$ candidates. Let $D^*$ denote the front-end winner with retrieval confidence $S^*$. Rather than relying on pointwise binary relevance decisions, DEC treats $D^*$ as the reference and learns the relative preference of each challenger by jointly calibrating their continuous visual support and retrieval confidence.

For each candidate document $D_i$, a concise visual-attribute description $VD_i$ is extracted from visually descriptive sections identified by the document catalog. The VLM calibrator takes the query image $I$, the target entity derived from $Q$, the title of $D_i$, and $VD_i$ as input. The continuous visual-support signal is computed as the logit difference:

$$Z_i = z_i^{\text{sup}} - z_i^{\text{rej}} = \log \frac{p_{\text{cal}}(\text{sup} \mid I, Q, D_i, VD_i)}{p_{\text{cal}}(\text{rej} \mid I, Q, D_i, VD_i)}. \tag{10}$$

where $z_i^{\text{sup}}$ and $z_i^{\text{rej}}$ denote the corresponding calibrator logits. The continuous score $Z_i$ preserves differences in support strength for subsequent cross-candidate comparison. Let $Z^*$ denote the visual-support score of $D^*$, with its calibration score fixed as $H^* = 0$. For each challenger $D_i$, we construct relative evidence with respect to $D^*$:

$$\Delta Z_i = Z_i - Z^*, \qquad \Delta \log S_i = \log \frac{S_i}{S^*}, \tag{11}$$

$$H_i = w_z \Delta Z_i + w_s \Delta \log S_i + b, \qquad w_z, w_s \geq 0, \tag{12}$$

where $w_z$, $w_s$, and $b$ are learned gate parameters. Among $D^*$ and all challengers, the document with the highest calibration score is selected as the final Top-1 document $\widehat{D}$, followed by lightweight text section reranking for answer generation.

### 4.4 TRAINING OBJECTIVES

VAR adopts a multi-stage training strategy with the vision encoder and knowledge-base image index frozen. Each stage uses a separately constructed training subset and is optimized after freezing all preceding components.

**Stage 1. Visual Anchor Selector.** The selector is trained with complementary ranking and embedding supervision. Ranking targets are derived from the gold-document rank of each anchor-conditioned query over the full knowledge base. The embedding target is the most similar image associated with the gold document under the best-performing anchor. Ranking supervision uses softmax cross-entropy, while embedding supervision uses multi-positive InfoNCE to impose a differentiable supervision signal on the anchor.

**Stage 2. Multi-Anchor Scorer.** With the selector frozen, the second stage learns candidate-level ranking over retrieved documents, which form a listwise training group. Listwise softmax cross-entropy is used to rank the gold document above competing candidates.

**Stage 3. VLM Calibrator.** The third stage introduces description-grounded supervision over the Top-$K_{CA}$ candidates. The gold document is labeled Support and the others are labeled Reject; if absent, all candidates are labeled Reject. The calibrator combines support–reject cross-entropy with listwise and margin objectives to strengthen support for the correct candidate.

**Stage 4. Gate.** The gate is trained with listwise softmax cross-entropy to assign the highest calibrated score to the gold document.

Detailed objectives are provided in Appendix B.2.

## 5 EXPERIMENTS

### 5.1 DATASETS AND METRICS

VAR is evaluated on three KB-VQA benchmarks: **InfoSeek** (Chen et al., 2023), **E-VQA** (Mensink et al., 2023), and **OK-VQA** (Marino et al., 2019). InfoSeek comprises 1.3M VQA pairs over 11K

Table 2: **Retrieval performance on E-VQA and InfoSeek.** "w/o Calibration" and "w. Calibration" denote VAR without and with the DEC stage, respectively. † Metrics that remain identical w/o and w. calibration, since DEC only reranks candidates within the Top-$K_{CA}$ set. Best results are in bold.

| Method | E-VQA | | | | InfoSeek | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@20 | R@1 | R@5 | R@10 | R@20 |
| *Global Visual Evidence Retrieval* | | | | | | | | |
| CLIP I-T (Radford et al., 2021) | 3.3 | 7.7 | 12.1 | 16.5 | 32.0 | 54.0 | 61.6 | 68.2 |
| Wiki-LLaVA (Caffagni et al., 2024) | 3.3 | – | 9.9 | 13.2 | 36.9 | – | 66.1 | 71.9 |
| LLM-RA (Jian et al., 2024) | – | – | – | – | 47.3 | 53.8 | – | – |
| *Multimodal Evidence Retrieval* | | | | | | | | |
| $mR^2AG$ (Zhang et al., 2026) | – | – | – | – | 38.0 | – | 65.0 | 71.0 |
| ReflectiVA (Cocchi et al., 2025) | 15.6 | 36.1 | – | 49.8 | 56.1 | 77.6 | – | 86.4 |
| EchoSight (Yan & Xie, 2024) | 36.5 | 47.9 | 48.8 | 48.8 | 53.2 | 74.0 | 77.4 | 77.9 |
| Wiki-R1 (Ning et al., 2026) | 16.7 | 41.0 | 44.8 | 47.5 | 46.9 | 67.1 | 72.9 | 77.2 |
| KBMR (Xu et al., 2026) | 24.7 | 48.7 | 52.4 | 55.0 | 60.3 | 74.7 | 79.6 | 84.4 |
| *Fine-Grained Visual Evidence Retrieval* | | | | | | | | |
| OMGM (Yang et al., 2025a) | 42.8 | 55.7 | 58.1 | 58.7 | 64.0 | 80.8 | 83.6 | 84.8 |
| Region-R1 (Hu & Tu, 2026) | 44.7 | 48.2 | 49.9 | 50.8 | 66.5 | 78.2 | 80.2 | 81.7 |
| **VAR (Ours)** *w/o Calibration* | 42.0 | 61.1 | **67.8** | **69.9** | 61.2 | 80.7 | **84.4** | **87.5** |
| **VAR (Ours)** *w. Calibration* ($K_{CA}$=6) | **45.6** | **61.4** | **67.8**† | **69.9**† | **68.5** | **80.9** | **84.4**† | **87.5**† |

OVEN images; following the standard protocol, retrieval uses a 100K Wikipedia knowledge base and evaluation is performed on the full validation set. E-VQA contains over 221K question–answer pairs from iNaturalist and Google Landmarks v2, with a WikiWeb2M-derived knowledge base of about 2M Wikipedia articles; results are reported on the official test set. OK-VQA contains about 14K knowledge-intensive questions over COCO images, using the E-VQA Wikipedia knowledge base for retrieval. Retrieval performance is assessed via Recall@K, measuring whether the ground-truth document appears in the Top-K results. Answer accuracy follows the official metric of each benchmark: standard VQA accuracy for InfoSeek and OK-VQA, and BEM score for E-VQA.

### 5.2 Implementation Details

EVA-CLIP-8B serves as the frozen vision encoder, with all knowledge-base images pre-encoded and indexed by FAISS-GPU. Visual anchors are constructed from high-level tokens between the fifth-to-last and fourth-to-last Transformer layers. VAM evaluates 64 candidate 3×3 token groups on a 16×16 grid, selects the Top-$K_{VA}$ = 8 anchors, and re-encodes them through the remaining 4 frozen layers. Each anchor-conditioned query retrieves the Top-$K_D$ = 50 documents, whose scores are integrated by a two-layer multi-anchor scorer with the frozen EchoSight (Yan & Xie, 2024) Q-Former score using a weight of 0.5. The resulting Top-$K_{CA}$ = 6 candidates, comprising the original Top-1 and five challengers, are passed to DEC. The visual anchor selector and multi-anchor scorer are trained on 6K and 28K examples, respectively. The VLM Calibrator uses 6K examples, followed by a 1K split for gate calibration. The calibrator is built on Qwen2.5-VL-3B-Instruct and fine-tuned with LoRA. For answer generation, a fine-tuned MiniLM selects the Top-3 question-relevant sections from the final Top-1 document. Data construction and optimization hyperparameters are detailed in the Appendix. All experiments use four NVIDIA A800-SXM4-80GB GPUs and take approximately 6 hours in total.

### 5.3 Main Results

**Retrieval Results.** Table 2 compares VAR with representative retrieval methods on E-VQA and InfoSeek. VAR achieves the best performance on both benchmarks. VAM substantially improves R@5–20, indicating that visual anchors recover gold document that is missed by the global representation rather than merely reshuffling existing candidates. DEC further lifts R@1 to 45.6 on E-VQA and 68.5 on InfoSeek, showing that DEC effectively converts the candidate coverage gained by VAM into correct Top-1 decisions by resolving ambiguity among already-retrieved entities.

**Answer Generation Results.** Table 3 shows that VAR achieves the best reported answer accuracy without task-specific generator fine tuning, reaching 50.6/45.4 on E-VQA and 48.1/46.6/47.0 on

Table 3: **Answer accuracy comparison on E-VQA and InfoSeek.** Gen. FT indicates task-specific fine-tuning of the answer generator. Best results are in bold.

| Method | Model | Gen. FT | Retriever | E-VQA | | InfoSeek | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Single-Hop | All | Unseen-Q | Unseen-E | All |
| *Zero-shot MLLMs* | | | | | | | | |
| BLIP-2 (Li et al., 2023) | Flan-T5$_{XL}$ | – | – | 12.6 | 12.4 | 12.7 | 12.3 | 12.5 |
| GPT-4V (OpenAI, 2023) | – | – | – | 26.9 | 28.1 | 15.0 | 13.3 | 14.6 |
| Qwen2.5-VL-7B (Bai et al., 2025) | – | – | – | 21.7 | 20.3 | 22.8 | 24.1 | 23.7 |
| *Retrieval-Augmented Models* | | | | | | | | |
| EchoSight (Yan & Xie, 2024) | Mistral-7B/LLaMA-3-8B | × | EVA-CLIP-8B | 41.8 | - | - | - | 31.3 |
| OMGM (Yang et al., 2025a) | LLaMA-3-8B | × | EVA-CLIP-8B | 49.9 | – | 35.3 | 33.6 | 34.4 |
| CC-VQA (Hong et al., 2026) | Qwen2.5-VL-7B | × | EVA-CLIP-8B | 41.4 | 36.1 | 44.7 | 46.1 | 45.1 |
| ReflectiVA (Cocchi et al., 2025) | LLaMA-3.1-8B | ✓ | EVA-CLIP-8B | 28.0 | 29.2 | 40.4 | 39.8 | 40.1 |
| MMKB-RAG (Ling et al., 2025) | Qwen2-VL-7B-Instruct | ✓ | EVA-CLIP-8B | 39.7 | 35.9 | 36.4 | 36.3 | 36.4 |
| KBMR(MMKB-RAG) (Xu et al., 2026) | Qwen2-VL-7B-Instruct | ✓ | KBMR | 47.3 | 43.1 | 45.6 | 45.7 | 45.7 |
| Wiki-PRF (Hong et al., 2025) | Qwen2.5-VL-7B | ✓ | EVA-CLIP-8B | 37.1 | 36.0 | 43.3 | 42.7 | 42.8 |
| Wiki-R1 (Ning et al., 2026) | Qwen2.5-VL-7B | ✓ | EVA-CLIP-8B + ColBERTv2 | 41.0 | 37.1 | 47.8 | 42.3 | 44.1 |
| **VAR (Ours)** | Qwen2.5-VL-7B-Instruct | × | EVA-CLIP-8B | 44.8 | 42.0 | 45.5 | 45.4 | 45.5 |
| **VAR (Ours)** | LLaMA-3.1-8B-Instruct | × | EVA-CLIP-8B | **50.6** | **45.4** | 46.1 | 46.5 | 46.4 |
| **VAR (Ours)** | Qwen2.5-7B-Instruct | × | EVA-CLIP-8B | 45.4 | 42.1 | **48.1** | **46.6** | **47.0** |

Table 4: **Performance on OK-VQA.** Qw.VL. denotes Qwen2.5-VL-7B-Instruct.

| Method | Model | Gen. FT | Accuracy |
|---|---|---|---|
| Qw.VL. | Qw.VL. | × | 60.7 |
| OMGM | LLaMA-3-8B | × | 66.6 |
| MMKB-RAG | Qw.VL. | ✓ | 65.4 |
| CC-VQA | Qw.VL. | × | 78.8 |
| **VAR (Ours)** | Qw.VL. | ✓ | **80.6** |

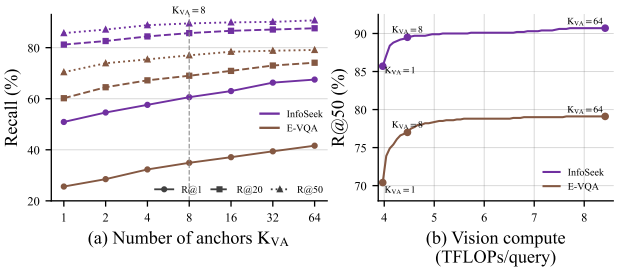



Figure 3: **Visual Anchor Budget Trade-off.**

InfoSeek. The gains are consistent across generators, with LLaMA-3.1-8B-Instruct performing best on E-VQA and Qwen2.5-7B-Instruct on InfoSeek, indicating that VAR's retrieval improvements transfer reliably to downstream answering. Text-only LLMs outperform the VLM, likely because answering mainly relies on retrieved textual evidence. Additional visual input may be under-utilized or introduce cross-modal interference (Tong et al., 2024; Liu et al., 2026).

On OK-VQA, we fine-tune Qwen2.5-VL-7B-Instruct on the training split to accommodate domain shift. VAR again outperforms the compared KB-VQA methods (Table 4), demonstrating effectiveness under a substantially different knowledge-intensive VQA setting.

### 5.4 Ablation Study

**Contribution of Core Components.** Table 5 evaluates the contributions of VAM and DEC. The baseline performs global EVA-CLIP retrieval. Adding VAM substantially improves retrieval recall on both datasets, showing that visual anchors recover complementary gold-document evidence underrepresented in the global representation. Building on VAM, DEC further improves R@1/QA by 7.3/3.9 points on InfoSeek and 3.6/4.7 points on E-VQA, demonstrating its effectiveness in correcting Top-ranked candidates and yielding the best overall performance.

**Generalization to Unseen Entities.** Although VAM mines visual anchors without question guidance, it generalizes well to unseen entities. On the Unseen Entity split, VAR achieves **59.7/79.8/84.3/87.4** R@1/5/10/20, closely matching the full-set results of **61.2/80.7/84.4/87.5** in Table 5. This suggests that retrieval supervision already provides a useful signal for highlighting transferable cues without explicit question conditioning.

**Ablation on the Number of Selected Visual Anchors ($K_{VA}$).** Figure 3(a) shows that recall increases with more anchors but gradually saturates, indicating complementary yet diminishing retrieval gains. At R@50, Figure 3 (b) shows that Top-8 retains 98.7%/97.4% of All-64 performance on InfoSeek/E-VQA while reducing anchor-dependent replay computation and retrieval cache by

Table 5: **Component ablation of VAR.**

| Components | | InfoSeek | | | | | E-VQA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VAM | DEC | R@1 | R@5 | R@10 | R@20 | QA | R@1 | R@5 | R@10 | R@20 | QA |
| × | × | 45.2 | 68.7 | 74.8 | 79.9 | 34.2 | 14.7 | 32.8 | 42.0 | 50.4 | 23.9 |
| ✓ | × | 61.2 | 80.7 | 84.4 | 87.5 | 43.1 | 42.0 | 61.1 | 67.8 | 69.9 | 45.9 |
| × | ✓ | 57.1 | 68.9 | 74.8 | 79.9 | 40.1 | 23.9 | 33.8 | 42.0 | 50.4 | 31.6 |
| ✓ | ✓ | **68.5** | **80.9** | **84.4** | **87.5** | **47.0** | **45.6** | **61.4** | **67.8** | **69.9** | **50.6** |

Table 6: **Analysis of DEC.**

(a) **Relative Calibration Effect on R@1.**

| Method | InfoSeek | E-VQA |
|---|---|---|
| Uncalibrated | 63.3 | 35.8 |
| Calibrated | **68.5** | **45.6** |

(b) **Learned gate parameters.**

| Dataset | $w_z$ | $w_s$ | $b$ |
|---|---|---|---|
| InfoSeek | 0.48 | 9.69 | 0.04 |
| E-VQA | 1.21 | 12.70 | 0.05 |

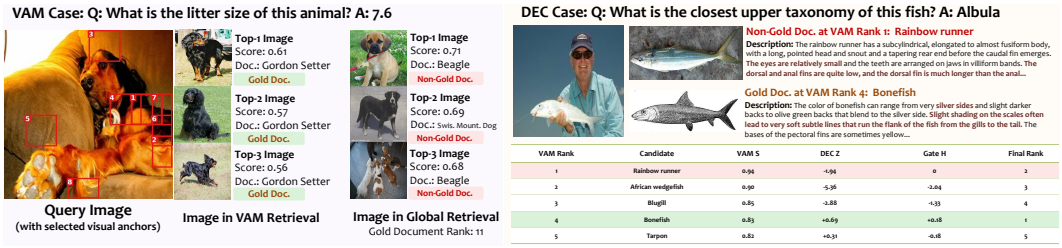



Figure 4: **Case studies of VAM and DEC.**

87.5%. Including fixed full-image encoding, total vision-side computation decreases from 8.419 to 4.464 TFLOPs. We therefore set $K_{VA} = 8$ as the default coverage–efficiency trade-off.

**Effect of Relative Calibration in DEC.** We examine whether independently estimated visual-support scores are sufficient for Top-1 document selection. *Uncalibrated* directly ranks the Top-$K_{CA}$ candidates by their support–reject margins $Z_i$. *Calibrated*, corresponding to the full DEC, instead learns each challenger's relative preference with respect to the front-end winner. As shown in Table 6(a), directly ranking candidates by $Z_i$ achieves 63.3/35.8 R@1 on InfoSeek/E-VQA, whereas relative calibration improves the results to 68.5/45.6. The substantial gains indicate that pointwise visual-support estimation alone does not sufficiently resolve competition among plausible candidates. Explicitly learning their relative preferences provides a stronger signal for Top-1 document selection. Table 6(b) further explains the learned gate. Although $w_s > w_z$, $\Delta \log S$ has a much smaller raw scale than $\Delta Z$, making the weighted terms comparable. Besides, since lower-ranked candidates typically have $\Delta \log S_i < 0$, a challenger must provide sufficiently stronger visual support to offset its retrieval deficit, yielding conservative Top-1 correction.

**Case Studies.** Figure 4 illustrates the complementary roles of VAM and DEC. In the VAM case, global retrieval confuses the *Gordon Setter* with similar breeds and ranks the gold document 11th. VAM instead mines complementary anchors over the head, coat, body, and contextual regions, allowing multiple anchor-conditioned queries to consistently retrieve evidence for the gold entity. In the DEC case, VAM retrieves *Bonefish* at Rank 4 but ranks the visually similar *Rainbow runner* first. DEC assigns stronger relative continuous visual support to *Bonefish*, and the gate promotes it to Rank 1. Together, VAM improves fine-grained candidate retrieval, while DEC resolves residual ambiguity through relative comparison among high-ranked candidates.

## 6 Conclusion

We reveal an overlooked phenomenon in KB-VQA retrieval: entity-discriminative information can persist in high-level vision tokens yet remain underexpressed in global retrieval embeddings. Our **VAR** framework mines these **Visual Anchors** to guide `[CLS]` updates through the remaining frozen encoder layers. It further uses description-grounded continuous visual support to learn relative preferences among high-ranked candidates, jointly calibrating visual support and retrieval confidence for Top-1 document selection. Experiments demonstrate state-of-the-art entity retrieval on InfoSeek and E-VQA and answer accuracy on these benchmarks and OK-VQA. A limitation is that anchor selection is currently not question-guided; incorporating question semantics may enable more task-aware visual evidence acquisition.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and code debugging during method implementation. We did not use generative AI tools to develop theoretical or conceptual frameworks, formulate mathematical claims or proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, generate or process research data, conduct qualitative analysis, or interpret experimental results; the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to survey and summarise related literature and to assist with manuscript writing, editing, and formatting. We have reviewed all AI-assisted work: LLM-assisted code was verified and tested for correctness by three authors, and AI-assisted text was repeatedly and critically reviewed and revised. The core methodology and innovations were developed by the authors based on our own experimental observations, and the use of generative AI did not affect the scientific rigor or originality of the research. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

The dataset utilised in this study is primarily curated from publicly available open-source datasets. Furthermore, the interaction scenarios are strictly confined to routine daily assistance. Consequently, this work does not raise any ethical concerns.

### REPRODUCIBILITY STATEMENT

To support reproducibility, we provide comprehensive details on our data sources, the data construction pipeline, model hyperparameters, and the specific prompts used throughout the process. Upon acceptance, we will release additional reproducibility resources and further refine our materials based on reviewer feedback.

## REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 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*, pp. 14948–14968, 2023.

Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9199–9209, 2025.

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

Yuyang Hong, Jiaqi Gu, Qi Yang, Lubin Fan, Yue Wu, Ying Wang, Kun Ding, Shiming Xiang, and Jieping Ye. Knowledge-based visual question answer with multimodal processing, retrieval and filtering. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yuyang Hong, Jiaqi Gu, Yujin Lou, Lubin Fan, Qi Yang, Ying Wang, Kun Ding, Yue Wu, Shiming Xiang, and Jieping Ye. Cc-vqa: Conflict- and correlation-aware method for mitigating knowledge conflict in knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5232–5241, 2026.

Chan-Wei Hu and Zhengzhong Tu. Region-r1: Reinforcing query-side region cropping for multimodal re-ranking. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10492–10505, 2026.

Pu Jian, Donglei Yu, and Jiajun Zhang. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10939–10956, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742, 2023.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 5294–5316, 2024.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zihan Ling, Zhiyao Guo, Yixuan Huang, Yi An, Shuai Xiao, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. Mmkb-rag: A multi-modal knowledge-based retrieval-augmented generation framework. *arXiv preprint arXiv:2504.10074*, 2025.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, 2026.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6417–6431, 2021.

Qian Ma, SM Rayeed, Charles V Stewart, Qiong Wu, and Yao Ma. Identifying and resolving pitfalls of knowledge-based vqa benchmarks: Auditing, repairing, and augmenting. *arXiv preprint arXiv:2607.00159*, 2026a.

Qian Ma, Qiong Wu, Zhengyi Zhou, and Yao Ma. Ground then rank: Revisiting knowledge-based vqa with training-free entity identification. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11617–11635, 2026b.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, 2019.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14111–14121, 2021.

GuangHao Meng, Sunan He, Jinpeng Wang, Tao Dai, Letian Zhang, Jieming Zhu, Qing Li, Gang Wang, Rui Zhang, and Yong Jiang. Evdclip: Improving vision-language retrieval with entity visual descriptions from large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6126–6134, 2025.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3113–3124, 2023.

Shan Ning, Longtian Qiu, and Xuming He. Wiki-r1: Incentivizing multimodal reasoning for knowledge-based vqa via data and sampling curriculum. In *International Conference on Learning Representations*, 2026.

OpenAI. GPT-4V(ision) System Card, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40: 2413–2427, 2018.

Jialin Wu and Raymond Mooney. Entity-focused dense passage retrieval for outside-knowledge visual question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8061–8072, 2022.

Hangrui Xu, Zhengxian Wu, Yunyao Yu, Zhuohong Chen, Rui Cong, Xiangwen Deng, Zhifang Liu, Peng Jiao, and Haoqian Wang. Beyond visual similarity: Entity-aligned retrieval for knowledge-based visual question answering. *arXiv preprint arXiv:2608.21450*, 2026.

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

Wei Yang, Jingjing Fu, Rui Wang, Jinyu Wang, Lei Song, and Jiang Bian. Omgm: Orchestrate multiple granularities and modalities for efficient multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24545–24563, 2025a.

Zhen Yang, Zhuo Tao, Qi Chen, Liang Li, Yuankai Qi, Anton van den Hengel, and Qingming Huang. Separation of powers: On segregating knowledge from observation in llm-enabled knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24753–24762, 2025b.

Wei Ye, Yixin Su, Yueguo Chen, Longxiang Gao, Jianjun Li, Ruixuan Li, and Rui Zhang. QKVQA: Question-focused filtering for knowledge-based VQA, 2026.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Auto-controlled image perception in mllms via visual perception tokens. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21822–21831, 2025.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, Jin Ma, Ying Shan, and Weiming Hu. mr$^2$ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *IEEE Transactions on Multimedia*, 2026.

This appendix provides supplementary materials and results. First, Appendix A details the prompts used in the experiments. Next, Appendix B and Appendix C present experimental details and additional experiments. Finally, Appendix D shows qualitative results.

# A PROMPTS

## A.1 PROMPT FOR DESCRIPTION-GROUNDED EVIDENCE CALIBRATION

**System Prompt**

You are a visual entity compatibility verifier for Wikipedia retrieval. Use the attached query image as the primary evidence. The document description may contain non-visual facts, so judge only whether it is visually compatible with the depicted target entity. Output exactly A when the candidate document is visually supported, or B when it is not supported. Output one symbol and nothing else.

**User Prompt**

Target entity: {target_entity} // Extracted from the question text using simple rules.
Candidate title: {candidate_title}
Visual description: {candidate_visual_description}
// Visual descriptions are rule-extracted from sections indicated by headings such as "Feature" or
//"Description"; otherwise, the opening paragraph is used as a fallback.
Is this candidate document visually compatible with the depicted target entity? Decision:

## A.2 PROMPTS FOR ANSWER GENERATION

**System Prompt for Infoseek**

Answer encyclopedic questions using the provided context. Return only the concise answer. Use numerical form for quantities, dates, or times when applicable.

**System Prompt for E-VQA**

Answer the encyclopedic question using the provided context. If the question requires multiple distinct answers, return all answers supported by the context as a comma-separated list; otherwise, return a single concise answer. Use numerical form for quantities, dates, or times when applicable. Return only the answer, without explanations or numbering.

**System Prompt for OK-VQA**

Answer the question using the provided context. Return only a short and concise answer, preferably a single word or phrase. Do not provide explanations or additional text.

**User Prompt for Infoseek/E-VQA/OK-VQA**

Context: {context}
Question: {question}
Answer:

# B EXPERIMENT DETAILS

## B.1 RETRIEVAL FEATURES $x_{i,v}$ AND $\gamma_i$

$x_{i,v} \in \mathbb{R}^5$ is a per-anchor retrieval feature vector describing the retrieval state of candidate document $D_i$ under the anchor-conditioned representation $q_v$.

| Feature | Dim. Index | Definition |
|---|---|---|
| Retrieval presence | 1 | Whether $D_i$ appears in the Top-$K_D$ retrieval list under $q_v$ |
| Reciprocal rank | 2 | $1/r_{i,v}$ |
| Retrieval similarity | 3 | $q_v^\top e_{i,v}^*$ |
| Score margin | 4 | Difference from the Top-1 retrieval score |
| Selector confidence | 5 | Selector probability of anchor $v$ |

$\gamma_i \in \mathbb{R}^4$ is a cross-anchor aggregation feature vector summarizing how consistently and strongly the selected anchors support candidate $D_i$.

| Feature | Dim. Index | Definition |
|---|---|---|
| Retrieval frequency | 1 | Fraction of selected anchors retrieving $D_i$ |
| Maximum retrieval score | 2 | Highest similarity of $D_i$ across supporting anchors |
| Mean retrieval score | 3 | Mean similarity over supporting anchors |
| Spatial support span | 4 | Maximum distance between centers of supporting anchors |

### B.2 Training Objectives

We detail the four-stage objectives introduced in Section 4.4. The vision encoder, knowledge-base image index, and all preceding components remain frozen in each stage.

**Stage 1. Visual Anchor Selector.** An offline evaluation of all candidate anchors provides the full-corpus gold-document ranks $r_{i,j}^{\text{gold}}$ of their anchor-conditioned queries. The embedding target $z_i^{\text{gold}}$ is the fixed normalized embedding of the gold-document image best matched by the query with the lowest gold-document rank.

**Ranking loss.** Given selector scores $\boldsymbol{s}_i = [s_{i,1}, \ldots, s_{i,J}]$, we construct soft ranking targets and minimize cross-entropy:

$$u_{i,j} = \frac{1}{\log_2(r_{i,j}^{\text{gold}} + 1)}, \qquad y_{i,j} = \frac{\exp(u_{i,j}/\tau_r)}{\sum_{k=1}^{J} \exp(u_{i,k}/\tau_r)}, \tag{13}$$

$$\mathcal{L}_{\text{rank}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{J} y_{i,j} \log \text{softmax}(\boldsymbol{s}_i)_j, \tag{14}$$

where $B$ is the global batch size.

**Embedding loss.** To make anchor selection differentiable during training, we compute soft anchor weights and aggregate them over each patch token:

$$a_{i,j} = \frac{\exp(s_{i,j}/\tau_s)}{\sum_{k=1}^{J} \exp(s_{i,k}/\tau_s)}, \qquad g_{i,p} = \sum_{j:\, x_{i,p} \in G_{i,j}} a_{i,j}. \tag{15}$$

At each replayed layer, these weights modify the `[CLS]`-to-patch attention. Only `[CLS]` is updated, while the layer-wise patch states cached from the natural forward pass are reused:

$$\widetilde{\alpha}_{i,p} = \mu_i \frac{\alpha_{i,p} g_{i,p}}{\sum_{r=1}^{N} \alpha_{i,r} g_{i,r}}, \qquad \widetilde{q}_i = \text{Norm}(f_{\text{tail}}(c_i, X_i; \widetilde{\boldsymbol{\alpha}}_i)), \tag{16}$$

where $\mu_i = \sum_{p=1}^{N} \alpha_{i,p}$ and Norm denotes $\ell_2$ normalization. This soft relaxation is used only during selector training; inference retains the hard Top-$K_{\text{VA}}$ anchors.

Let $D_i^{\text{gold}}$ and $I_i^{\text{gold}}$ denote the gold document and selected supporting image. Samples sharing either target form the positive set for multi-positive InfoNCE:

$$\mathcal{P}(i) = \left\{ k \in \{1, \ldots, B\} \,\middle|\, D_k^{\text{gold}} = D_i^{\text{gold}} \;\vee\; I_k^{\text{gold}} = I_i^{\text{gold}} \right\}, \tag{17}$$

$$\mathcal{L}_{\text{aln}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\sum_{k \in \mathcal{P}(i)} \exp(\widetilde{q}_i^\top z_k^{\text{gold}}/\tau_e)}{\sum_{k=1}^{B} \exp(\widetilde{q}_i^\top z_k^{\text{gold}}/\tau_e)}. \tag{18}$$

The selector objective is

$$\mathcal{L}_{\text{selector}} = \mathcal{L}_{\text{aln}} + \lambda \mathcal{L}_{\text{rank}}, \tag{19}$$

where $\tau_r$, $\tau_s$, and $\tau_e$ are temperature parameters, and $\lambda$ weights the ranking loss.

**Stage 2. Multi-Anchor Scorer.** The frozen selector produces candidate groups, of which those containing the gold document are used for training. With $S_i$ defined as the normalized candidate confidence in Eq. 9, the listwise cross-entropy is

$$\mathcal{L}_{\text{M}} = -\log S^{\text{gold}}, \tag{20}$$

where $S^{\text{gold}}$ is the confidence assigned to the gold document.

**Stage 3. VLM Calibrator.** For the Top-$\text{K}_{\text{CA}}$ candidates produced by the frozen front-end, the gold document is labeled $\text{sup}$ and all other candidates are labeled $\text{rej}$. Given candidate labels $t_i$, the classification loss is

$$\mathcal{L}_{\text{cls}} = -\frac{1}{\text{K}_{\text{CA}}} \sum_{i=1}^{\text{K}_{\text{CA}}} \log p_{\text{cal}}(t_i \mid I, Q, D_i, VD_i). \tag{21}$$

For groups containing the gold document, we additionally apply listwise and margin losses to the support logits:

$$\mathcal{L}_{\text{list}} = -\log \frac{\exp(z^{\text{sup}}_{\text{gold}})}{\sum_{i=1}^{\text{K}_{\text{CA}}} \exp(z^{\text{sup}}_i)}, \qquad \mathcal{L}_{\text{margin}} = \left[ m - z^{\text{sup}}_{\text{gold}} + \max_{i \neq \text{gold}} z^{\text{sup}}_i \right]_+, \tag{22}$$

where $z^{\text{sup}}_{\text{gold}}$ is the gold document's support logit, $m > 0$ is the margin, and $[x]_+ = \max(x, 0)$. The combined objective is

$$\mathcal{L}_{\text{cal}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{list}} \mathcal{L}_{\text{list}} + \lambda_{\text{margin}} \mathcal{L}_{\text{margin}}. \tag{23}$$

If the gold document is absent, all candidates are labeled $\text{rej}$ and only $\mathcal{L}_{\text{cls}}$ is applied.

**Stage 4. Gate.** The gate is trained on Top-$\text{K}_{\text{CA}}$ groups containing the gold document. Let $H_i$ denote the calibration score defined in Eq. 12, with $H_i = H^* = 0$ when $D_i = D^*$. The listwise objective is

$$\mathcal{L}_{\text{gate}} = -\log \frac{\exp(H_{\text{gold}})}{\sum_{i=1}^{\text{K}_{\text{CA}}} \exp(H_i)}, \qquad w_z, w_s \geq 0, \tag{24}$$

where $H_{\text{gold}}$ is the calibration score of the gold document.

### B.3 Implementation Details

Table 7 summarizes the training subsets. Tables 8 and 9 detail the selector and scorer architectures, respectively. LN and GN denote layer normalization and group normalization.

**Stage 1: Visual Anchor Selector.** We use AdamW with learning rate $10^{-4}$, weight decay $0.01$, global batch size 64, gradient clipping at 1.0, and dropout 0.1. The InfoNCE temperature is $\tau_e$ = 0.07, and the ranking-loss weight in Eq. 19 is $\lambda$ = 0.1.

**Stage 2: Multi-Anchor Scorer.** Training uses batch size 8, weight decay $10^{-4}$, gradient clipping at 5.0, and dropout $0.1$.

**Stage 3: VLM Calibrator.** Qwen2.5-VL-3B-Instruct is adapted using LoRA with rank $r$ = 64, scaling parameter $\alpha$ = 128, and dropout 0.05, applied to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. We use AdamW with learning rate $5 \times 10^{-6}$, weight decay $0.01$, gradient accumulation over 4 steps, and gradient clipping at 1.0. Gradient checkpointing and FP16 mixed-precision training are enabled. In Eq. 23, $\lambda_{\text{list}}$ = 1.0 and $\lambda_{\text{margin}}$ = 0.5; the margin in Eq. 22 is $m$ = 0.5.

**Stage 4: Gate.** L-BFGS-B optimizes the gate parameters subject to $w_z, w_s \geq 0$. No additional regularization or absolute-support feature is used.

Table 7: **Training-data composition for the four stages.**

| Stage | Size | Data Construction |
|---|---|---|
| Visual Anchor Selector | 6K | Difficulty-balanced sampling: 45% Top-1 preserve (global Top-1 hits); 15% Rank-1 rescue (global Top-1 misses recovered by at least one anchor); 40% harder cases (neither global retrieval nor any anchor ranks the gold document first). |
| Multi-Anchor Scorer | 28K | Entity-balanced sampling over 3,500 gold entities, averaging 8 questions per entity with a maximum of 12. Candidate groups are generated by the frozen selector. |
| VLM Calibrator | 6K | Independent samples processed by the frozen front-end to construct Top-$K_{CA} = 6$ candidate groups. |
| Gate | 1K | Disjoint from the VLM Calibrator subset. Candidate groups are constructed as above, retaining only those containing the gold document. |

Table 8: **Architecture of the Visual Anchor Selector.** The selector scores 64 candidate anchors from 256 high-level patch tokens of dimension 4096. $B$ denotes batch size. The frozen vision encoder is excluded from the parameter count.

| Component | Operation | Input $\rightarrow$ Output | Params. |
|---|---|---|---|
| Patch projection 1 | Linear + LN + GELU + Dropout | $4096 \rightarrow 2048$ | 8,394,752 |
| Patch projection 2 | Linear + LN + GELU + Dropout | $2048 \rightarrow 1024$ | 2,100,224 |
| Patch projection 3 | Linear + LN + GELU + Dropout | $1024 \rightarrow 512$ | 525,824 |
| Anchor extraction | 64 spatial $3 \times 3$ token groups | $[B, 256, 512] \rightarrow [64B, 512, 3, 3]$ | 0 |
| Spatial scorer 1 | $1 \times 1$ Conv + GN + GELU | $512 \rightarrow 256$ | 131,840 |
| Spatial scorer 2 | $3 \times 3$ Conv + GN + GELU + residual | $256 \rightarrow 256$ | 590,592 |
| Spatial scorer 3 | $1 \times 1$ Conv + GN + GELU | $256 \rightarrow 64$ | 16,576 |
| Spatial pooling | $3 \times 3$ mean pooling | $[64B, 64, 3, 3] \rightarrow [64B, 64]$ | 0 |
| Score head 1 | Linear + GELU + Dropout | $64 \rightarrow 32$ | 2,080 |
| Score head 2 | Linear | $32 \rightarrow 1$ | 33 |
| **Total** | | | **11,761,921** |

# C Additional Experiments

## C.1 Training Dynamics of the VLM Calibrator

Figure 5 presents the optimization dynamics of the VLM calibrator. The overall training loss decreases steadily, and the support–reject classification, listwise, and margin losses exhibit similar downward trends. This behavior indicates stable joint optimization of candidate discrimination and relative visual-support estimation. The validation loss reaches its minimum value of 0.982 at step 300 and increases thereafter, while the training loss continues to decrease. This divergence suggests the onset of overfitting. We therefore select the checkpoint at step 300 for subsequent evaluation.

## C.2 Inference Time

Table 10 compares the inference latency of VAR with representative KB-VQA methods (Yang et al., 2025a; Ye et al., 2026; Hong et al., 2025; 2026; Hu & Tu, 2026). OMGM and QKVQA introduce additional retrieval, reranking, and candidate filtering stages, while Wiki-PRF and CC-VQA further rely on costly visual processing or reasoning modules. Region-R1 reports 2.61 s for bounding-box generation and region cropping alone, rather than the complete pipeline.

VAR requires 1.61 s per sample, including 1.32 s before answer generation. Its visual-anchor mining reuses high-level features from the frozen vision encoder and preserves the original knowledge-base index, while description-grounded calibration is restricted to a small candidate set. This avoids repeated visual encoding and large-scale knowledge processing, keeping the additional online compu-

Table 9: **Architecture of the Multi-Anchor Scorer.** The scorer processes eight anchor-conditioned queries. Query and knowledge-base image embeddings are 1280-dimensional. $G$ denotes the number of candidate documents.

| Component | Operation | Input → Output | Params. |
|---|---|---|---|
| Pairwise interaction | Concatenate $q \odot e,\ \lvert q-e\rvert,\ q,\ e$ | $[G, 8, 4\times 1280] \to [G, 8, 5120]$ | 0 |
| Interaction encoder | Linear, no bias | $5120 \to 64$ | 327,680 |
| Retrieval-feature encoder ($x_{i,v}$) | Linear | $5 \to 64$ | 384 |
| Feature normalization | LN | $[G, 8, 64] \to [G, 8, 64]$ | 128 |
| Anchor interaction ($\times 2$) | Transformer, 4 heads, FFN 128 | $[G, 8, 64] \to [G, 8, 64]$ | 66,944 |
| Attention pooling | Linear + Softmax + weighted sum | $[G, 8, 64] \to [G, 64]$ | 65 |
| Anchor mean pooling | Mean over anchors | $[G, 8, 64] \to [G, 64]$ | 0 |
| Anchor max pooling | Max over anchors | $[G, 8, 64] \to [G, 64]$ | 0 |
| Pooling concatenation | Attention / Mean / Max | $3\times 64 \to 192$ | 0 |
| Cross-anchor encoder ($\gamma_i$) | Linear + LN + GELU | $4 \to 64$ | 448 |
| Final concatenation | Pooled + cross-anchor features | $192 + 64 \to 256$ | 0 |
| Scorer 1 | Linear + LN + GELU + Dropout | $256 \to 128$ | 33,152 |
| Scorer 2 | Linear + GELU + Dropout | $128 \to 64$ | 8,256 |
| Scorer 3 | Linear | $64 \to 1$ | 65 |
| **Total** | | | **437,122** |

The interaction encoder implements the four block-wise projections in Eq. 8 through a single linear map over concatenated features. The retrieval features $x_{i,v}$ and $\gamma_i$ are defined in Appendix B.1.

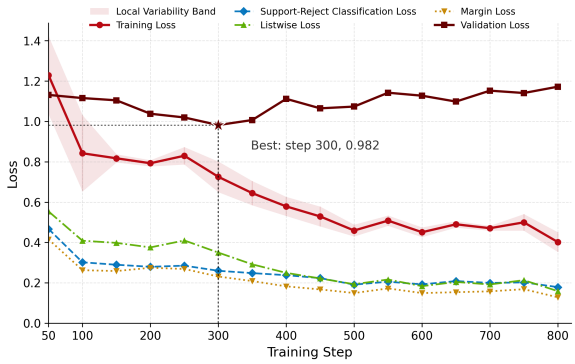



Figure 5: **Training dynamics of the VLM calibrator.**

Table 10: **Inference efficiency comparison.** Latency is reported in seconds per sample.

| Method | GPU | Pre-Generation | Generation | Total |
|---|---|---|---|---|
| **VAR** | A800 | **1.32** | **0.29** | **1.61** |
| OMGM | L40 | 2.16 | 0.39 | 2.55 |
| QKVQA | L40 | 2.39 | 0.38 | 2.77 |
| Wiki-PRF | A800 | 7.90 | 0.74 | 8.77[†] |
| CC-VQA | A800 | 4.99 | 3.95 | 8.94 |
| Region-R1 | A6000 | 2.61[‡] | – | – |

[†]The reported total is 8.77 s, while the listed stages sum to 8.64 s. [‡]Bounding-box generation and cropping only. "–" denotes unavailable timing.

tation lightweight. Since hardware and timing protocols differ across methods, these results should be interpreted as practical latency references rather than a strictly controlled speed comparison.

### C.3 Analysis of Anchor-Conditioned Replay Depth.

We conduct an Oracle study on 2K InfoSeek samples to investigate how the intervention depth of Visual Anchors affects their retrieval potential. As in our main method, we apply the same anchor mechanism to patch tokens cached from different depths of the 32-block EVA-CLIP vision encoder. These tokens come from the natural forward pass and retain the interactions accumulated up to the corresponding layer. The anchor changes how the CLS token accesses these cached states, after which the CLS representation is updated through the remaining frozen Transformer blocks.

As shown in Table 11, Oracle Best-of-64 R@1 remains remarkably stable from After Block 1 through After Block 28, varying only from 68.25 to 68.75. This consistency also holds at the sample level: After Block 1 and After Block 20 agree on the Rank-1 hit status for 99.25% of the samples, and 99.93% of their jointly successful cases share at least one Rank-1 spatial window. These results suggest that the spatial supports associated with retrieval-relevant evidence remain highly stable across a broad range of depths.

Table 11: **Analysis of anchor-conditioned replay depth.** Oracle Best-of-64 R@1 measures whether any of the 64 spatial supports can produce a Rank-1 retrieval result when the anchor intervention is initiated at different encoder depths.

| Intervention Position | Remaining Blocks | Oracle R@1 | Replay Latency (ms) |
|---|---|---|---|
| After Block 1 | 31 | 68.40 | $303.45 \pm 81.94$ |
| After Block 4 | 28 | 68.35 | $243.65 \pm 5.73$ |
| After Block 8 | 24 | 68.45 | $207.68 \pm 1.56$ |
| After Block 12 | 20 | 68.60 | $173.38 \pm 0.87$ |
| After Block 16 | 16 | 68.65 | $138.78 \pm 0.72$ |
| **After Block 20** | 12 | **68.75** | $104.19 \pm 0.59$ |
| After Block 24 | 8 | 68.65 | $69.56 \pm 0.45$ |
| After Block 28 (ours) | 4 | 68.25 | $34.99 \pm 0.28$ |
| After Block 30 | 2 | 66.15 | $17.71 \pm 0.15$ |
| After Block 31 | 1 | 59.30 | $9.30 \pm 3.39$ |

Importantly, this does not imply that early-layer tokens already constitute complete semantic Visual Anchors. Instead, earlier intervention leaves more frozen Transformer blocks through which the focused CLS can repeatedly aggregate and reorganize increasingly high-level semantic evidence from the cached natural-forward patch states. Thus, early intervention mainly provides more opportunities for evidence integration before forming the final retrieval representation.

This interpretation is further supported near the end of the encoder. Oracle R@1 decreases from 68.25 after Block 28 to 66.15 after Block 30, and then sharply to 59.30 after Block 31. Therefore, identifying a useful spatial support is not sufficient by itself; adequate remaining depth is still required to integrate the selected evidence into a global representation compatible with the pretrained retrieval space. The remaining Transformer blocks thus provide **evidence reassembly capacity** after the spatial intervention.

This additional capacity, however, comes with rapidly increasing computation. Replay after Block 20 requires 104.19 ms under the same 64-window protocol, whereas replay after Block 28 requires only 34.99 ms. We therefore choose Block 28 as a lightweight operating point: the visual representation has already undergone extensive global contextualization, while only the final 4 Transformer blocks need to be replayed. This reduces both inference-time overhead and training-time re-encoding cost while retaining substantial retrieval potential.

### C.4 The Nature of Visual Anchors: Evidence Reorganization within a Shared Semantic Space

The analysis of anchor-conditioned replay depth also provides insight into the nature of Visual Anchors. Rather than merely modifying an attention map, Visual Anchors perform **internal visual evidence reorganization**: they revisit high-level representations already encoded by the vision encoder, identify evidence relevant to the current task, and recompose it through a frozen replay path. This reconstruction remains within the pretrained semantic space, allowing the resulting representation to emphasize a different subset of evidence while remaining compatible and directly comparable with the original global representation.

This perspective motivates a broader interpretation: **same knowledge, different tasks, different evidence routes**. An intuitive analogy is learning, taking notes, and then sitting an examination. An examination determines which parts of one's knowledge need to be expressed in the final answer, but the answer itself does not exhaust everything that was learned or recorded in the notes. For example, a high-school mathematics olympiad and a university mathematics final examination may both involve overlapping knowledge of calculus, yet emphasize different concepts and require different reasoning patterns. When the task changes, it may therefore be unnecessary to relearn the underlying material. Instead, one can merely revisit the same notes, retrieve a different subset of relevant knowledge, and reorganize it into a new answer.

From a broader perspective, this suggests a research direction beyond the KB-VQA setting. A pretrained model may already encode useful evidence for multiple downstream tasks, even when

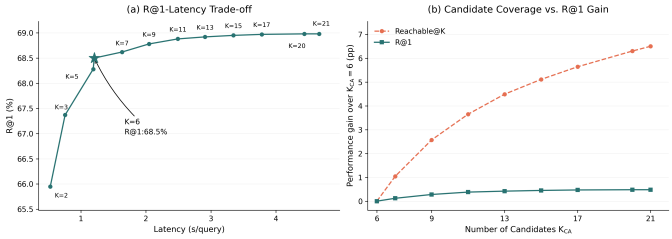



Figure 6: **Ablation on the calibration candidate depth $\mathrm{K_{CA}}$.** (a) Final R@1 versus calibration latency under different candidate depths. (b) Additional candidate coverage and the corresponding R@1 gain relative to $\mathrm{K_{CA}} = 6$.

Table 12: **Ablation on the EchoSight Q-Former fusion weight $\alpha$ on InfoSeek.**

| $\alpha$ | R@1 | R@5 | R@10 | R@20 |
|---|---|---|---|---|
| 0.00 | 58.4 | 78.7 | 82.9 | 86.4 |
| 0.25 | 60.6 | 80.1 | 83.9 | 87.2 |
| **0.50** | **61.2** | **80.7** | **84.4** | **87.5** |
| 0.75 | 53.9 | 77.2 | 82.4 | 86.7 |
| 1.00 | 17.1 | 46.8 | 62.3 | 77.4 |

such evidence is not explicitly emphasized by its default output interface. Future models could therefore learn not only how to encode information, but also how to recognize what they have already encoded and autonomously construct different evidence routes according to new task demands. Such a mechanism would allow a shared pretrained representation to support different tasks by selectively reorganizing existing internal evidence, rather than repeatedly relearning the underlying input.

### C.5 Ablation on Q-Former Fusion Weight

We further study the fusion weight of the EchoSight Q-Former score. As shown in Table 12, adding Q-Former evidence improves R@1 from 58.4 at $\alpha = 0$ to 61.2 at $\alpha = 0.5$. This improvement may arise from the complementary information captured by the two signals: VAM emphasizes fine-grained visual evidence within the vision encoder, whereas the Q-Former additionally aligns visual representations with section-level text. Notably, the fused score outperforms either signal used in isolation, indicating that the two components capture complementary retrieval evidence rather than redundant cues. Performance drops as either source becomes dominant, with the largest degradation observed when relying solely on the Q-Former score. These results support a balanced fusion, and we therefore set $\alpha = 0.5$ for all experiments.

### C.6 Ablation on the Calibration Candidate Depth $\mathrm{K_{CA}}$

Figure 6 studies the number of candidate documents sent to the calibrator. Increasing $\mathrm{K_{CA}}$ initially brings clear gains: from 5 to 6 candidates, R@1 improves from 68.10 to 68.47, while latency increases only marginally from 1.19 s to 1.21 s. Beyond this point, however, the return rapidly diminishes. When $\mathrm{K_{CA}}$ increases from 6 to 21, Reachable@K rises substantially from 81.23 to 87.73, whereas final R@1 improves by only 0.48 points (68.47→68.95), with latency increasing from 1.21 s to 4.67 s. This growing gap between candidate coverage and realized Top-1 gain indicates that candidate availability is no longer the main bottleneck once a moderate candidate set is reached. We therefore set $\mathrm{K_{CA}} = 6$, consisting of the original Top-1 document and five challengers, as the default accuracy–efficiency trade-off.

## D QUALITATIVE RESULTS

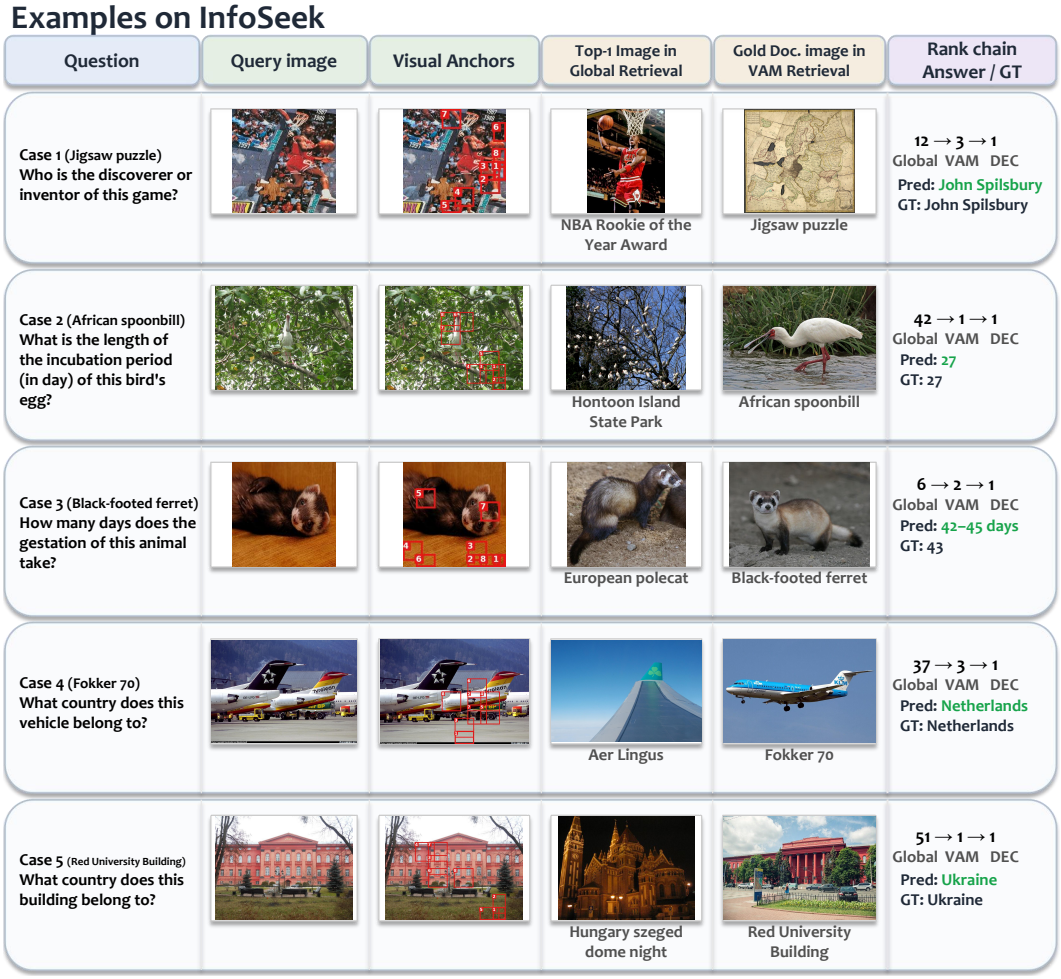



Figure 7: **Qualitative examples on InfoSeek.** Each row shows the question, query image, selected visual anchors, the Top-1 image from global retrieval, the knowledge-base image associated with the gold document, its rank progression through VAM and DEC, and the final prediction together with the ground-truth answer.

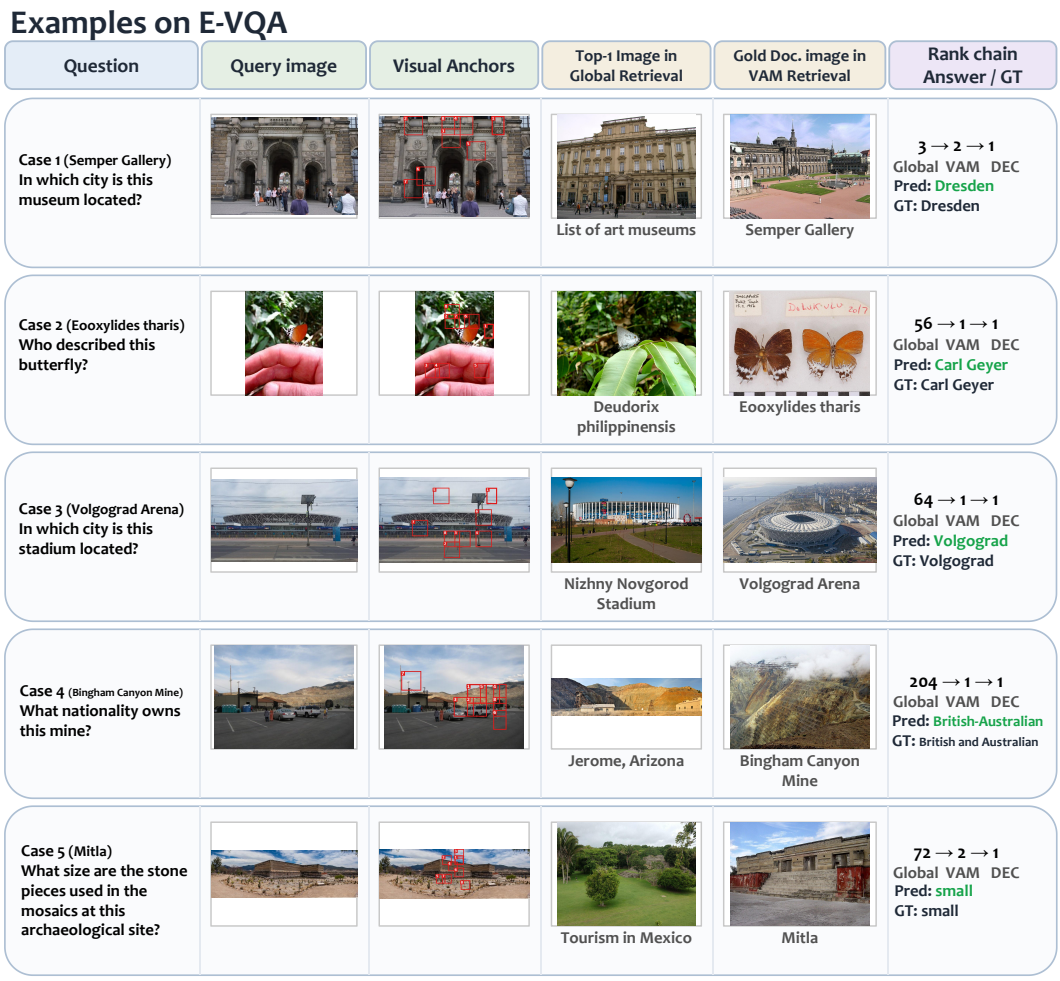



Figure 8: **Qualitative examples on E-VQA.** Each row shows the question, query image, selected visual anchors, the Top-1 image from global retrieval, the knowledge-base image associated with the gold document, its rank progression through VAM and DEC, and the final prediction together with the ground-truth answer.

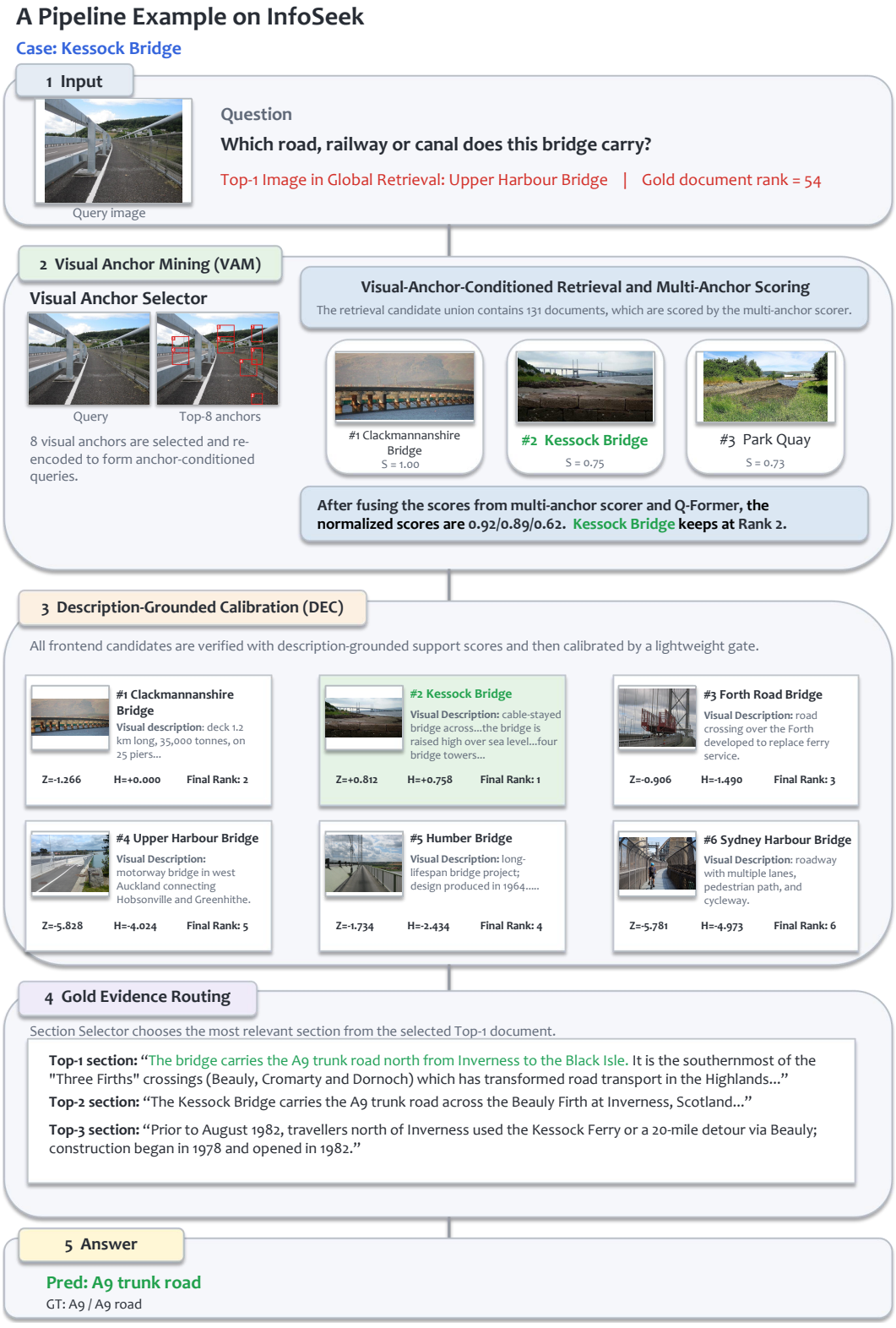



Figure 9: **End-to-end VAR pipeline visualization on InfoSeek.**




Figure 10: **End-to-end VAR pipeline visualization on E-VQA.**