# CAER: CONFLICT-AWARE EVIDENCE REFINEMENT FOR KNOWLEDGE-BASED VISUAL QUESTION ANSWERING

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

## ABSTRACT

Knowledge-based Visual Question Answering (KB-VQA) requires models to jointly understand visual content and retrieve external knowledge for factual reasoning. Existing retrieval-augmented methods typically follow a retrieve-then-generate paradigm, while recent training-free approaches further address conflicts between retrieved knowledge and parametric knowledge through conflict-aware decoding. However, these methods primarily treat knowledge conflict as generation-stage uncertainty, without exploiting it to refine retrieved evidence. Consequently, noisy knowledge and entity mismatch may still propagate into subsequent reasoning, limiting the effective utilization of external knowledge and degrading factual accuracy. In this paper, we propose **CAER** (**C**onflict-**A**ware **E**vidence **R**efinement), a training-free framework for KB-VQA that leverages knowledge conflict as feedback for evidence refinement. Specifically, CAER first performs entity-aware evidence acquisition to alleviate entity mismatch among retrieved candidates. It then conducts conflict-aware evidence refinement within a fixed candidate pool by exploiting detected conflicts to optimize evidence selection. Finally, CAER estimates evidence reliability by jointly modeling question relevance, entity consistency, visual alignment, and conflict consistency, and propagates the resulting reliability signal to a trust-guided knowledge fusion module that dynamically balances retrieved evidence and parametric knowledge during answer generation. On InfoSeek, CAER achieves 48.2% accuracy while remaining fully training-free, outperforming the strongest training-free baseline by 3.58 percentage points. These results demonstrate the effectiveness of conflict-aware evidence refinement for improving factual reasoning in knowledge-based visual question answering.

## 1 INTRODUCTION

Knowledge-Based Visual Question Answering (KB-VQA) lies at the intersection of visual understanding, language reasoning, and external knowledge retrieval. Given an image and a question, a KB-VQA system must not only recognize visual content but also retrieve relevant facts from external knowledge sources (e.g., Wikipedia) to generate reliable answers. Although vision-language models (VLMs) have demonstrated remarkable visual reasoning capabilities, their parametric knowledge is fixed after pre-training and often insufficient for knowledge-intensive queries. Consequently, external knowledge retrieval remains indispensable for factual reasoning.

Despite its success, retrieval augmentation does not necessarily improve reasoning reliability. Existing KB-VQA methods generally follow a retrieval-then-generation paradigm, where retrieved knowledge is directly injected into a VLM for answer generation. However, the retrieved evidence itself is often unreliable. Visual retrieval may introduce entity drift by retrieving visually similar yet semantically different entities, while even entity-aligned evidence may contain conflicting facts across different sources or disagree with the model's parametric knowledge. As a result, unreliable evidence can propagate throughout the reasoning process and ultimately degrade answer quality.

Recent training-free KB-VQA methods attempt to mitigate this issue by handling knowledge conflicts during decoding. Representative methods, including CoCoAKhandelwal et al. (2025), AdaCADWang et al. (2025), and CC-VQAHong et al. (2026), maintain multiple knowledge streams and dynamically



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.

regulate their contributions according to uncertainty- or consistency-related signals. Although these methods have demonstrated promising results, there remains room for improving how retrieved evidence is optimized and utilized during reasoning.

First, retrieved evidence is rarely optimized before reasoning. Existing methods primarily exploit conflict signals during answer generation, after the retrieved evidence has already been incorporated into the reasoning context. Consequently, entity-mismatched or conflicting evidence remains in the reasoning process, since conflict information cannot be propagated back to improve evidence quality before decoding.

Second, existing methods generally rely on prediction uncertainty rather than explicitly modeling evidence reliability. They typically estimate the contribution of different knowledge sources using uncertainty-related signals, such as entropy, output correlation, or confidence discrepancies. Although these signals reflect the uncertainty of model predictions, they do not explicitly assess whether retrieved evidence is consistently supported by the question, visual observations, and entity consistency. As a result, unreliable evidence may still exert excessive influence on the final prediction.

Motivated by these observations, we propose CAER (Conflict-Aware Evidence Refinement for Knowledge-Based Visual Question Answering), a training-free framework for conflict-aware evidence optimization and reliability-guided knowledge fusion. Rather than treating knowledge conflict solely as a decoding-stage uncertainty, CAER leverages conflict as an actionable signal for improving retrieved evidence before reasoning.

Specifically, CAER establishes a conflict-aware evidence-refinement feedback paradigm and comprises three stages. First, Entity-aware Evidence Acquisition incorporates the visual entity predicted by a VLM as an auxiliary semantic signal to rerank visually retrieved candidates and alleviate entity drift. Second, Conflict-aware Evidence Refinement and Reliability Modeling propagates detected conflict signals back to the evidence layer, refines candidate evidence before final answer generation, and estimates sentence-level reliability by jointly considering question relevance, entity consistency, visual alignment, and conflict consistency. Third, Reliability-guided Knowledge Fusion incorporates the estimated evidence reliability into Trust-based Conflict-Aware Fusion (TCAF), dynamically balancing external evidence and the VLM's parametric knowledge during decoding.

CAER is fully training-free and requires neither parameter optimization nor task-specific fine-tuning. It operates solely with frozen VLMs and external knowledge sources during inference.

The main contributions of this work are summarized as follows:

- We propose CAER, a training-free conflict-aware evidence refinement framework that treats knowledge conflict not merely as decoding-stage uncertainty, but as an actionable signal for evidence optimization.
- We develop a unified three-stage framework comprising entity-aware evidence acquisition, conflict-aware evidence refinement and reliability modeling, and reliability-guided knowledge fusion, enabling conflict signals to improve both evidence quality and knowledge utilization.
- Extensive experiments on InfoSeek demonstrate that CAER achieves the best performance among existing training-free KB-VQA methods without parameter updates.

## 2 Related Work

### 2.1 Retrieval-Augmented KB-VQA

Knowledge-Based Visual Question Answering (KB-VQA) requires models to jointly reason over visual content and external knowledge. To support knowledge-intensive reasoning, most existing methods adopt a retrieval-augmented pipeline that retrieves external knowledge before answer generation. EchoSight Yan & Xie (2024) improves retrieval quality through multimodal reranking over EVA-CLIPSun et al. (2023) retrieved evidence. Wiki-LLaVACaffagni et al. (2024) and MMKB-RAGLing et al. (2025) enhance retrieval with hierarchical retrieval strategies and multimodal knowledge bases. ReflectiVACocchi et al. (2025) and Wiki-PRFHong et al. (2025) instead improve knowledge utilization through reflective reasoning and reinforcement learning. More recently, ProMSAWu et al. (2026)

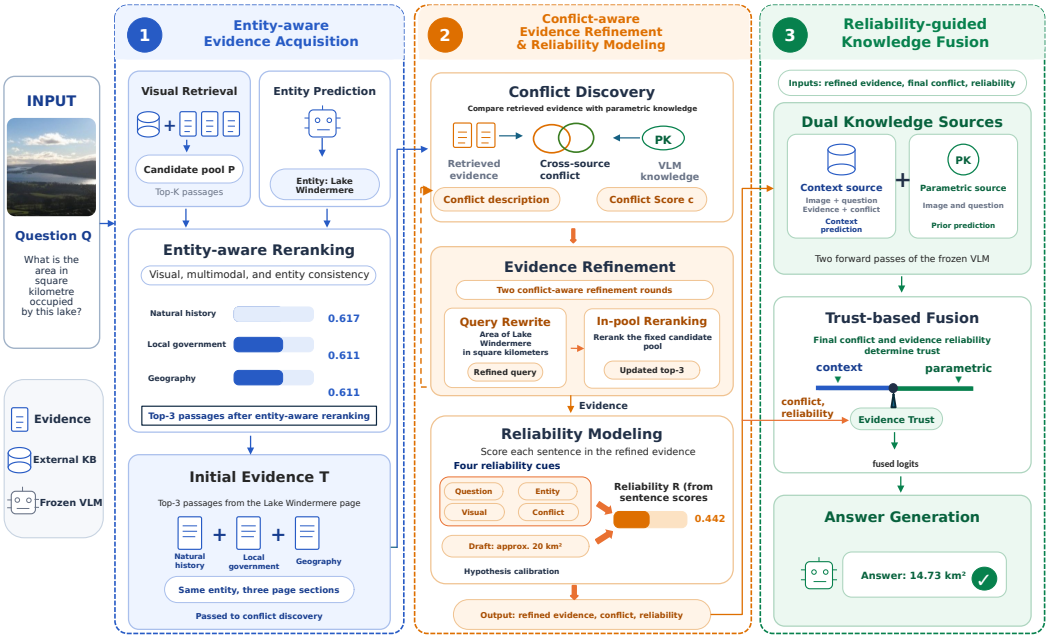


Figure 1: Overview of CAER. CAER consists of three stages: (1) Entity-aware Evidence Acquisition, (2) Conflict-aware Evidence Refinement and Reliability Modeling, and (3) Reliability-guided Knowledge Fusion.

formulates KB-VQA as an agent-based retrieval problem by learning to dynamically select retrieval tools. These methods improve different components of retrieval-augmented KB-VQA, including retrieval, reranking, and knowledge utilization. In contrast, our work focuses on improving the quality of retrieved evidence during inference without additional training.

### 2.2 Conflict-Aware Knowledge Reasoning

As retrieval augmentation becomes increasingly prevalent, resolving conflicts between retrieved evidence and parametric knowledge has attracted growing attention. Prompt-based methods, such as FaithfulRAG Zhang et al. (2025) and Context-Faithful Prompting Zhou et al. (2023), improve contextual faithfulness through prompt engineering and self-reflection. Training-free decoding methods instead regulate answer generation using confidence-related signals. CADShi et al. (2024) contrasts decoding distributions with and without retrieved contexts, AdaCADWang et al. (2025) dynamically fuses multiple decoding distributions using Jensen–Shannon divergence, and CoCoA further incorporates entropy-based confidence estimation. In multimodal settings, CC-VQA combines visual conflict reasoning with relevance-guided encoding and decoding to alleviate cross-modal knowledge conflicts. These approaches mainly exploit conflict information during reasoning. In contrast, our method further propagates conflict signals to the evidence layer for evidence refinement and explicitly estimates evidence reliability to guide conflict-aware knowledge fusion.

## 3 Method

### 3.1 Overview

CAER (Conflict-Aware Evidence Refinement) is a training-free framework for knowledge-based visual question answering (KB-VQA). Given an image $I$ and a question $Q$, CAER exploits external knowledge while reducing the influence of irrelevant evidence, entity mismatch, and knowledge conflict. As illustrated in Fig. 1, CAER consists of three stages. First, **Entity-aware Evidence Acquisition** introduces entity consistency as an auxiliary signal to improve the quality of visually

retrieved candidates. Second, **Conflict-aware Evidence Refinement and Reliability Modeling** uses detected conflicts to refine candidate evidence and estimates its reliability from multiple consistency signals. Third, **Reliability-guided Knowledge Fusion** dynamically balances retrieved evidence and parametric knowledge according to the estimated evidence reliability. CAER is fully training-free and requires no additional parameter optimization or task-specific fine-tuning.

## 3.2 Entity-aware Evidence Acquisition

### 3.2.1 Candidate Retrieval and Entity-aware Reranking

In KB-VQA, knowledge retrieval requires not only finding visually relevant information but also ensuring that the retrieved knowledge corresponds to the correct visual entity. Existing retrieval methods usually rely on image-level semantic similarity: $s_{vis}(I, p)$, which effectively expands the candidate space but may introduce entity mismatch. For example, different animals, buildings, or landmarks may share similar visual appearances while corresponding to completely different factual knowledge.

To alleviate this issue, CAER does not modify the initial retrieval process. Instead, it introduces entity consistency as an additional signal for posterior correction within the retrieved candidate set.

Given an input image $I$, CAER first predicts the dominant visual entity:

$$e = VLM(I), \tag{1}$$

where $e$ denotes the predicted entity. Then, candidate passages are retrieved from the external knowledge base:

$$P = \{p_1, p_2, \ldots, p_K\}. \tag{2}$$

The candidate passages are obtained using an EVA-CLIP-based FAISS index. For each candidate passage $p_j$, CAER performs entity-aware reranking by combining visual relevance, multimodal relevance, and entity consistency:

$$S(p_j) = \alpha_1 s_{vis} + (1 - \alpha_1) s_{multi}, \tag{3}$$

where $s_{multi} = \frac{s_{mm}+s_{ent}}{2}$ , $\alpha_1$ controls the balance between visual retrieval similarity and multimodal reranking signals. Following EchoSight, we use $\alpha_1$ to balance the FAISS retrieval similarity ($s_{vis}$, from EVA-CLIP image–image matching) against multimodal reranking signals. Within the reranking term, BLIP2 multimodal relevance ($s_{mm}$, image–question–passage matching via Q-Former) and entity text consistency ($s_{ent}$, EVA-CLIP text–text matching) are combined with equal weight. The initial evidence set is obtained as: $T = \{p_1, \ldots, p_N\}$.

Note that the entity information is not used as a hard retrieval constraint. Instead, it serves as an auxiliary reranking signal, avoiding retrieval failures caused by entity recognition errors while reducing entity mismatch caused by visual ambiguity.

## 3.3 Conflict-aware Evidence Refinement and Reliability Modeling

### 3.3.1 Conflict Discovery and Evidence Refinement

Although entity-aware reranking improves evidence quality, retrieved passages may still contain inconsistent facts from different sources. Therefore, CAER introduces a conflict discovery module before answer generation.

Given image $I$, question $Q$, and retrieved evidence $T$, a VLM analyzes potential conflicts:

$$(R_{conf}, c) = \Phi(I, Q, T), \tag{4}$$

where $R_{conf}$ denotes the conflict description and $c$ represents the overall conflict score. $\Phi$ is implemented as a single VLM.

CAER considers several types of conflicts, including entity inconsistency, contradictory attributes, and semantic disagreement. When:$c > \theta$, CAER activates evidence refinement. Instead of directly suppressing conflicting knowledge during decoding, CAER uses conflict information as feedback to optimize evidence selection. Specifically, a refined query is generated:

$$Q' = VLM(Q, e, I, R_{conf}, \mathcal{H}), \tag{5}$$

where $I$ is the input image and $\mathcal{H}$ denotes the history of previous queries to avoid repetition. The VLM is prompted to generate a conflict-focused refinement query that targets the specific disputed fact in $R_{conf}$, anchored to the predicted entity $e$, while remaining distinct from previous queries in $\mathcal{H}$. Then, CAER reranks passages within the fixed candidate pool:

$$S'(p_j) = f(Q', p_j), \quad p_j \in P. \tag{6}$$

where $f$ is the same fusion scoring function as $S(\cdot)$ in Eq. (3), with the refined query $Q'$ replacing $Q$ in the question-dependent terms. No new retrieval request is issued; refinement only re-orders the existing pool. The refined evidence set $T^*$ consists of the top-$N$ passages of $P$ under $S'$.

#### 3.3.2 Evidence Reliability Modeling

After evidence refinement, CAER estimates the reliability of retrieved knowledge. Reliable evidence should be consistent with the question, target entity, visual content, and detected conflicts.

#### 3.3.3 Multi-signal Reliability Estimation

The refined evidence is decomposed into sentences $\mathcal{T}^* = \{s_1, \ldots, s_M\}$. For each sentence $s_i$, CAER computes four reliability signals: *question relevance* $r_q = \mathrm{sim}(Q', s_i)$, measuring whether the sentence addresses the refined question; *entity consistency* $r_e = \mathrm{sim}(e, s_i)$, measuring whether the sentence concerns the predicted visual entity rather than a visually similar one; *visual alignment* $r_v = \mathrm{sim}(I, s_i)$, measuring whether the sentence is compatible with the image content; and *conflict consistency* $r_c = 1 - c \cdot \mathrm{sim}(R_{\mathrm{conf}}, s_i)$, which discounts sentences that carry the disputed content: $\mathrm{sim}(R_{\mathrm{conf}}, s_i)$ measures the sentence's relevance to the conflict description, so that a sentence closely matching the conflict is penalized in proportion to the overall conflict intensity $c$, while sentences irrelevant to the dispute keep $r_c \approx 1$. Here $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity in the corresponding encoder space (text–text for $r_q, r_e, r_c$; image–text for $r_v$). The final sentence reliability is aggregated as:

$$\hat{s}_i = \alpha r_q + \beta r_e + \epsilon r_v + \gamma r_c, \tag{7}$$

where $\alpha, \beta, \epsilon, \gamma$ are weighting coefficients that control each signal's contribution.

#### 3.3.4 Hypothesis-guided Evidence Calibration

Retrieval-side signals alone may not fully reveal which evidence supports the final answer. Therefore, CAER introduces a lightweight hypothesis-guided calibration strategy. A draft answer is first generated:

$$a_{draft} = VLM(I, Q, T^*). \tag{8}$$

The entity and draft answer are combined as an answer hypothesis $h = e \oplus a_{\text{draft}}$, where $\oplus$ denotes plain text concatenation : the entity name followed by the draft answer. The sentence reliability is further adjusted:

$$\hat{s}_i \leftarrow \hat{s}_i + \delta \cdot sim(h, s_i), \tag{9}$$

where $\delta$ is a calibration coefficient. The draft answer is only treated as a hypothesis for identifying supporting evidence and is not directly used as the final prediction.

**Reliability Aggregation** The calibrated sentence scores are normalized into a distribution:

$$\pi_i = \frac{\exp(\hat{s}_i)}{\sum_{j=1}^{M} \exp(\hat{s}_j)}, \tag{10}$$

and the evidence concentration is measured by normalized entropy:

$$g = 1 - \frac{-\sum_{i=1}^{M} \pi_i \log \pi_i}{\log M}. \tag{11}$$

$g \in [0, 1]$ approaches 1 when the reliability mass concentrates on a few sentences and 0 when it is spread uniformly. Concentration alone, however, does not imply reliability, as the mass may concentrate on unreliable sentences. CAER therefore couples $g$ with the mean sentence reliability $\bar{s} = \frac{1}{M} \sum_i \hat{s}_i$, yielding a single *evidence reliability score*:

$$R = \bar{s} \cdot g, \tag{12}$$

which is passed to the fusion stage to modulate how strongly the context source is trusted.

### 3.4 Reliability-guided Knowledge Fusion

Existing conflict-aware methods mainly adjust decoding based on conflict or correlation signals. However, conflict intensity alone cannot fully determine whether external knowledge should be trusted. Therefore, CAER propagates the estimated evidence reliability to the final fusion stage, instantiated as Trust-based Conflict-aware Fusion (TCAF).

#### 3.4.1 Dual-source Knowledge Fusion

CAER considers two complementary knowledge sources: the context source, which conditions the VLM on the refined evidence and the conflict description,

$$(I, Q, T^*, R_{\text{conf}}) \rightarrow l_{ctx}(t),$$

and the parametric source, which relies only on the model's intrinsic knowledge,

$$(I, Q) \rightarrow l_{pri}(t),$$

where $l_{ctx}(t)$ and $l_{pri}(t)$ are the next-token logit vectors at decoding step $t$.

#### 3.4.2 Trust-based Conflict-aware Fusion

CAER estimates the trust score of external knowledge:

$$r_{trust} = \sigma\big(w_1(1 - c) + w_2 R\big), \tag{13}$$

where $\sigma$ is the sigmoid function; $(1 - c)$ rewards the absence of conflict, with $c$ re-estimated on the refined evidence after the final refinement round (the detection value from Eq. (4) is used when refinement is not triggered), so that successful refinement directly increases trust in the context source, and $w_1$ and $w_2$ are weighting coefficients. The final logits are fused as

$$l(t) = r_{trust}\, l_{ctx}(t) + (1 - r_{trust})\, l_{pri}(t), \tag{14}$$

where $l(t)$ is greedily decoded into the final answer. The trust weight $r_{trust}$ is computed once per sample and applied at every decoding step: since it is derived from the reliability of the evidence itself rather than from generation-time uncertainty, it does not fluctuate across tokens, which avoids the instability of token-adaptive weighting when early tokens are generic.

## 4 Experiment

### 4.1 Experimental Setup

**Dataset.** We evaluate CAER on InfoSeek Chen et al. (2023), a large-scale KB-VQA benchmark whose full split contains 1.35M question–answer pairs (934K train / 73K validation / 348K test) grounded in Wikipedia entities. Since the test-split answers are not publicly released, we follow the evaluation protocol of prior retrieval-augmented KB-VQA work Yan & Xie (2024); Hong et al. (2026) and report results on the filtered validation split, which contains 71,335 question–answer pairs over 36,655 images. Answering each question requires combining visual understanding with factual knowledge retrieved from an external Wikipedia corpus.

Table 1: Comparison with KB-VQA methods on InfoSeek. *Results copied from the original papers under their respective protocols (two metric columns merged). †Reproduced with the official code under our unified protocol. "VQA acc" is the official InfoSeek metric; "LLM judge" uses an identical judge for all unified-protocol rows. "Gen. FT" marks generation-model fine-tuning; EVA-CLIP denotes the 8B variant.

| | | | | InfoSeek (%) | |
|---|---|---|---|---|---|
| Method | Model | Gen. FT | Retriever | VQA acc | LLM judge |
| BLIP-2* Li et al. (2023) | Flan-T5$_{XL}$ | – | – | 12.5 | |
| InstructBLIP* Dai et al. (2023) | Flan-T5$_{XL}$ | – | – | 8.1 | |
| LLaVA-v1.5* Liu et al. (2024) | Vicuna-7B | – | – | 9.5 | |
| GPT-4V* Achiam et al. (2023) | – | – | – | 14.6 | |
| Qwen2.5-VL-7B* Bai et al. (2025) | Qwen2.5-VL-7B | – | – | 23.7 | |
| DPR$_{V+T}$* Lerner et al. (2024) | Multi-pass. BERT | – | CLIP-B/32 | 12.4 | |
| EchoSight* Yan & Xie (2024) | Mistral-7B/LLaMA-3-8B | ✗ | EVA-CLIP | 30.4 | |
| ReflectiVA* Cocchi et al. (2025) | LLaMA-3.1-8B | ✓ | EVA-CLIP | 40.1 | |
| MMKB-RAG* Ling et al. (2025) | Qwen2-7B | ✗ | EVA-CLIP | 36.4 | |
| Wiki-PRF* Hong et al. (2025) | Qwen2.5-VL-8B | ✓ | EVA-CLIP | 42.8 | |
| ProMSA* Wu et al. (2026) | Qwen2.5-VL-7B | ✓ | – | 49.2 | |
| EntRAG* Hu et al. (2026) | – | ✓ | – | 44.5 | |
| Qwen2.5-VL+RAG* Bai et al. (2025) | Qwen2.5-VL-7B | ✗ | EVA-CLIP | 41.8 | |
| *Unified protocol (identical backbone, retriever, knowledge base, and judge)* | | | | | |
| Qwen2.5-VL+RAG | Qwen2.5-VL-7B | ✗ | EVA-CLIP | 38.83 | 40.10 |
| CC-VQA† Hong et al. (2026) | Qwen2.5-VL-7B | ✗ | EVA-CLIP | 40.80 | 44.62 |
| **CAER (Ours)** | Qwen2.5-VL-7B | ✗ | EVA-CLIP | **43.90** | **48.20** |

**Knowledge Base and Retrieval.** Following previous retrieval-augmented KB-VQA methods, we use the official Wikipedia corpus as the external knowledge base. Images are encoded using EVA-CLIP and searched against a FAISS index to retrieve the top-K candidate passages according to visual similarity. Unlike conventional retrieve-then-generate pipelines, CAER performs entity-aware reranking and conflict-aware evidence refinement within the retrieved candidate pool without issuing additional retrieval requests.

**Implementation Details.** We adopt Qwen2.5-VL-7B as the frozen vision-language model throughout all experiments. No parameter updating or task-specific fine-tuning is involved during inference. All experiments are conducted using greedy decoding with the temperature set to zero. The retriever recalls the top-$K$=20 candidate passages via an EVA-CLIP-based FAISS index Johnson et al. (2021) (image–image matching), from which the top-$N$=3 passages after reranking are used for reasoning. Reranking uses BLIP-2 for multimodal scoring ($s_{mm}$) and EVA-CLIP for entity text matching ($s_{ent}$), with $\alpha_1$=0.5. Complete prompts, hyperparameter settings, and additional implementation details are provided in the supplementary material.

#### 4.1.1 Evaluation Metric

InfoSeek answers are open-ended entity mentions with highly variable surface forms, and exact string matching is known to be unreliable for answer equivalence Bulian et al. (2022). We therefore employ DeepSeek-V4-Pro (DeepSeek-AI et al., 2026) as the evaluator to compare the generated answer with the reference answer. The judge determines whether the prediction is semantically correct rather than requiring exact string matching, which is important for open-ended knowledge questions with multiple valid expressions. The judge temperature is set to 0 to ensure deterministic evaluation.

### 4.2 Main Results

Table 1 compares CAER with representative retrieval-augmented and training-free KB-VQA methods. Since prior results were obtained under heterogeneous backbones, knowledge bases, and evaluation protocols, we re-run the strongest training-free baseline, CC-VQA, under a unified protocol identical to ours (same backbone, retriever, knowledge base, and judge). Under this protocol, CAER achieves

Table 2: Ablation study on InfoSeek 10K subset.

| Configuration | Accuracy (%) | Δ |
|---|---|---|
| Full CAER | 46.42 | – |
| W/o Entity-aware | 45.72 | −0.70 |
| W/o Conflict Refinement | 45.46 | −0.96 |
| W/o Evidence Calibration | 45.90 | −0.52 |
| W/o Trust Fusion | 45.95 | −0.47 |
| W/o Reliability Estimation | 46.04 | −0.38 |
| W/ Generic Rewrite | 45.65 | −0.77 |
| W/ Generic Rewrite (uncond.) | 45.50 | −0.92 |

48.2% LLM-judge accuracy, outperforming CC-VQA$^{\dagger}$ by 3.58 points, and 43.9% official VQA accuracy, outperforming CC-VQA$^{\dagger}$ by 3.1 points; the consistent gain on both metrics indicates that the improvement is not an artifact of the LLM judge. Entries marked with $^{*}$ are copied from the original papers and are not strictly comparable; ProMSA and EntRAG, for instance, rely on task-specific fine-tuning and their own evaluation protocols. These results demonstrate that refining retrieved evidence according to detected knowledge conflicts is substantially more effective than handling conflicts solely during answer generation, without requiring any additional model training.

## 4.3 Ablation Study

### 4.3.1 Effectiveness of Individual Components

To verify the contribution of each proposed component, we conduct ablation experiments on a fixed 10K subset of the evaluation set, drawn by deterministic random sampling (seed 42; see supplementary material). The subset is used only for relative component comparisons. Each variant removes one module from the full system: **Full CAER**: All modules enabled (Entity-aware + Conflict Refinement + Reliability Modeling + Trust Fusion). **W/o Entity-aware**: Remove entity-aware reranking. **W/o Conflict Refinement**: Disable conflict-aware query rewriting and re-ranking. **W/ Generic Rewrite**: On the same triggered instances, keep the rewrite and reranking budget but omit the conflict description from the rewrite prompt. **W/ Generic Rewrite (uncond.)**: apply the generic rewrite to every question, regardless of the detected conflict. **W/o Evidence Calibration**: Remove hypothesis-guided evidence calibration. **W/o Trust Fusion**: Replace TCAF trust fusion with only the context stream logits. **W/o Reliability Estimator**: Replace the four-signal sentence scores by uniform weights and use a validation-fixed trust constant while retaining both decoding streams.

As shown in Table 2, removing any module degrades performance.Conflict-aware evidence refinement contributes the most ($-0.96$ points), followed by entity-aware acquisition ($-0.70$), hypothesis-guided calibration ($-0.52$), trust fusion ($-0.47$), and reliability estimation ($-0.38$).Two generic-rewrite controls attribute the refinement gain to the proposed mechanism rather than to the rewrite budget: rewriting every question with a generic prompt (uncond.) consumes more inference budget than triggered refinement yet recovers only 0.04 of the 0.96 points (45.50 vs. 45.46); restricting the same generic rewrite to triggered instances recovers 0.15 further points (45.65).

## 4.4 Qualitative Analysis

Figure 2 presents representative examples.

**Entity-aware reranking resolves visual ambiguity.** For Shwezigon Pagoda, the baseline's top-3 evidence is dominated by the visually similar *Bagan* region and answers *"Bagan"*. By incorporating the predicted entity as a soft reranking signal, CAER demotes the visually related but entity-mismatched passages and answers *"Nyaung-U, Myanmar"* correctly. This is the most frequent correction pattern: 6,459 of the 9,196 successful flips are attributed to entity-aware reranking.

**Conflict-aware refinement corrects unreliable evidence.** For **Lake Windermere**, CAER detects conflicting evidence ($c = 1.0$) and performs query refinement. The refined evidence contains the correct area value, leading to the answer *"14.73 km$^2$"* instead of the baseline prediction *"$\sim$56 km$^2$"*.

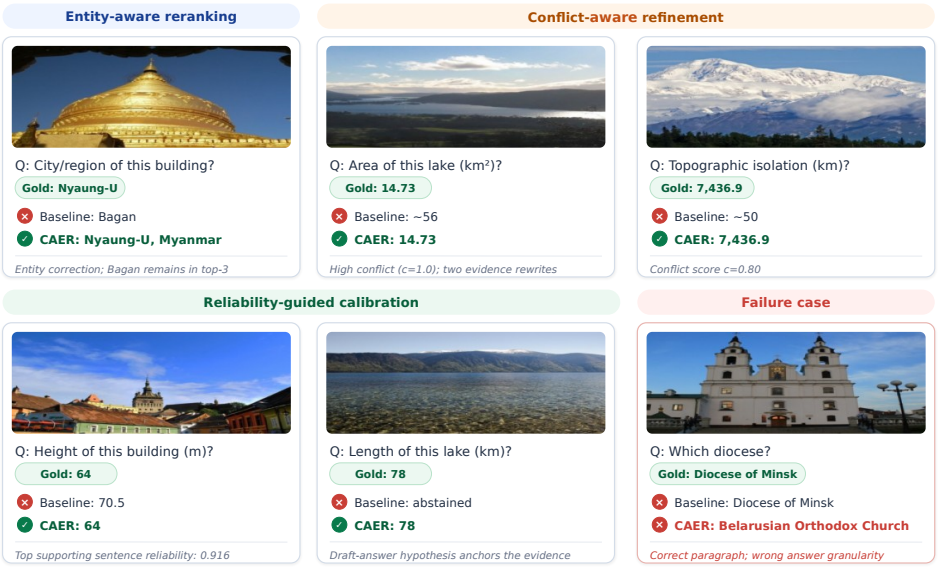


Figure 2: Qualitative examples on InfoSeek. Successful corrections are grouped by the component that drives them—entity-aware reranking, conflict-aware refinement, and reliability-guided calibration; the final case illustrates a typical failure caused by answer-granularity ambiguity.

Similarly, refinement improves the evidence quality for **Denali** by surfacing the passage containing the correct isolation value.

**Reliability-guided calibration improves factual selection.** For Sighisoara Clock Tower, the hypothesis-guided calibration identifies the sentence containing the correct height (*"64 meters"*), correcting the baseline error. For **Lake Teletskoye**, the entity-aware evidence ranking combined with reliability calibration selects the correct supporting knowledge among visually similar candidates.

**Failure case.** For Holy Spirit Cathedral in Minsk, CAER retrieves the correct evidence and even ranks the correct paragraph first, yet the calibration step extracts the answer at the wrong granularity (*"Belarusian Orthodox Church"* instead of *"Diocese of Minsk"*), while the baseline happens to answer correctly. This indicates that similarity-based calibration does not guarantee the correct answer granularity; failures also arise when early entity grounding is incorrect, in which case the error propagates to evidence acquisition.

## 5 Conclusion and Discussion

We have presented CAER, a training-free framework that introduces conflict-aware evidence refinement and reliability-guided knowledge fusion for KB-VQA. CAER first leverages entity-aware evidence acquisition to alleviate entity mismatch in visually retrieved candidates. Detected conflicts are then propagated back to the evidence layer to refine unreliable evidence before answer generation. Finally, evidence reliability is estimated from multiple consistency signals and incorporated into trust-based fusion to dynamically balance retrieved evidence and the VLM's parametric knowledge. Despite its effectiveness, CAER still has several limitations. First, reliability-based calibration may select incorrect answer granularity when evidence contains facts at multiple semantic levels, suggesting the need for answer-type-aware calibration. Second, CAER is evaluated on Wikipedia-grounded English KB-VQA benchmark with a 7B VLM backbone. Extending the framework to broader knowledge sources, larger models, and more complex multi-hop reasoning scenarios remains an interesting direction.

### AI USE STATEMENT

In this work, we used generative AI tools for language polishing and for drafting early versions of the illustrative figures, which were subsequently redrawn programmatically and verified by the authors. We did not use generative AI tools for research ideation, method design, experiment execution, or data analysis.

### ETHICS STATEMENT

This work uses only publicly available datasets (InfoSeek, OK-VQA) and a publicly available Wikipedia knowledge base; no new data involving human subjects was collected. The LLM judge is accessed via its official API in compliance with the provider's terms of use. The framework makes no demographic, biometric, or otherwise sensitive inferences, and we do not foresee harmful uses beyond those inherent to general-purpose question answering.

### REPRODUCIBILITY STATEMENT

CAER involves no training: all components are frozen off-the-shelf models used with deterministic decoding (temperature 0). The evaluation split, the 10K ablation subset (deterministic sampling, seed 42), all hyperparameters, and component versions are documented in Section 4.1 and Appendix B; the full judge prompt and scoring protocol are given in Appendix C.

## REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. Tomayto, tomahto. Beyond token-level answer equivalence for question answering evaluation. In *EMNLP*, pp. 291–305, 2022.

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 *CVPR*, pp. 13339–13348, 2024.

Yuxin Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Ming-Wei Chang, Kenton Lee, and Tania Bedrax-Weiss. Can pre-trained vision and language models answer visual information-seeking questions? In *EMNLP*, 2023.

Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. ReflectiVA: Enhancing reflective token-based augmentation in multimodal LLMs. In *CVPR*, pp. 24532–24543, 2025.

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

DeepSeek-AI et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

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 *NeurIPS*, 2025. Also known as Wiki-PRF / VLM-PRF, arXiv:2510.14605.

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 *CVPR*, 2026.

Yiheng Hu, Xiaoyang Wang, Qing Liu, Xiwei Xu, Qian Fu, et al. EntRAG: Entity-centric retrieval-augmented generation for knowledge-based visual question answering. In *ICML*, 2026.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

Anant Khandelwal, Manish Gupta, and Puneet Agrawal. Cocoa: Confidence-and context-aware adaptive decoding for resolving knowledge conflicts in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6846–6866, 2025.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, et al. Cross-modal retrieval for knowledge-based visual question answering. In *ECIR*, pp. 427–443, 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 *ICML*, 2023.

Yuping Ling, Runze Su, Zeyu Mi, Luoyu Li, Songmei Guo, Haoran Luo, et al. MMKB-RAG: A multi-modal knowledge-based retrieval-augmented generation framework. *arXiv preprint arXiv:2504.10074*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, pp. 8137–8146, 2019.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *NAACL*, pp. 783–791, 2024.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. AdaCAD: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *NAACL*, pp. 11636–11652, 2025.

ZhengXian Wu, Hangrui Xu, Kai Shi, Zhuohong Chen, Yunyao Yu, Chuanrui Zhang, Zirui Liao, Jun Yang, Zhenyu Yang, Haonan Lu, and Haoqian Wang. ProMSA: Progressive multimodal search agents for knowledge-based visual question answering. *arXiv preprint arXiv:2606.27974*, 2026.

Yibin Yan and Weidi Xie. EchoSight: Advancing visual-language models with wiki knowledge. In *Findings of EMNLP*, pp. 1538–1551, 2024.

Jiawei Zhang, Yu Zhao, Xiaohan Xu, Che Liu, Yaqi Su, Congying Han, Zhiqiang Xu, and Tiande Guo. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. *arXiv preprint arXiv:2506.08938*, 2025.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of EMNLP*, pp. 14544–14556, 2023.

# A Appendix

# B Additional Implementation Details

## B.1 Model and Retrieval Infrastructure

Table 3 lists the concrete components behind each module of CAER. All components are frozen off-the-shelf models; no parameter is updated at any stage.

Table 3: Components and versions. Decoding is greedy (temperature 0) throughout (§B.4); the judge is detailed in §C.

| Component | Configuration |
|---|---|
| Generator / analyzer $\Phi$ | Qwen2.5-VL-7B (frozen) |
| Entity prediction $e$ | same VLM, prompt $p_{\text{entity}}$ |
| Retrieval encoder | EVA-CLIP-8B, image–image |
| Vector index | FAISS-GPU 1.14.3, official Wikipedia corpus |
| Rerankers $s_{\text{mm}}$ / $s_{\text{ent}}$ | BLIP-2 Q-Former / EVA-CLIP text–text |

### B.2 Hyperparameters

Table 4 maps every coefficient appearing in the main paper's equations to its implementation value. $\alpha_1$ follows EchoSight; the remaining values were fixed on a validation probe before any test evaluation. The *Swept?* column states which knobs were varied: component on/off ablations and the aggregation strategy are reported in the main paper (Tables 2–3); continuous coefficients marked *single* were used at one configuration throughout. For the most sensitive knob, the conflict trigger $\theta$ that gates all evidence refinement, we swept $\theta \in \{0.3, 0.5, 0.7\}$ on the 10K ablation subset: accuracy is $46.0/46.1/46.4/45.9/45.46\%$ respectively, varying by at most $0.6$ pp and peaking at the adopted $\theta = 0.5$.

Table 4: Hyperparameters of CAER (equation numbers refer to the main paper). In *Swept?*: *single* = one configuration used throughout; *on/off* = ablated in the main paper (Tables 2–3); *fixed* = design constant.

| Symbol | Eq. | Meaning | Value | Swept? |
|---|---|---|---|---|
| $K$ / $N$ | (2) | candidate pool / evidence size | 20 / 3 | fixed |
| $\alpha_1$ | (3) | retrieval vs. rerank fusion | 0.5 | single |
| $\theta$ | — | conflict trigger ($c > \theta$) | 0.5 | {‘0, 0.3, 0.5, 0.7, 1.0} |
| $R_{\max}$ | — | max refinement rounds | 2 | on/off |
| $\alpha, \beta, \epsilon, \gamma$ | (7) | weights of $r_q, r_e, r_v, r_c$ | 0.5 / 0.3 / 0.15 / 0.2 | single |
| $\delta$ | (9) | hypothesis calibration | 0.2 | on/off |
| $w_2, w_3$ | (12) | trust fusion weights | 1.0 / 1.0 | fixed |
| aggregation | (10) | additive (beats geometric) | — | on/off |

### B.3 Evaluation Splits

Main results (main Table 1) use the full InfoSeek benchmark (71,335 question–answer pairs). Component ablations (main Tables 2–3) use a fixed 10K subset drawn by deterministic random sampling with seed 42.

### B.4 Runtime Environment

All experiments ran on 8× NVIDIA H20 (141 GB HBM3 each) under Ubuntu 22.04, with CUDA 13.0, PyTorch 2.11.0 + cu130, transformers 4.57.6, and FAISS-GPU 1.14.3. Decoding is greedy with temperature 0 and index search is exact.

### B.5 Prompts

All prompts below are reproduced verbatim from the code:

- $p_{\text{entity}}$ — visual entity prediction (Eq. 1);
- $\Phi$ — conflict analysis producing $(R_{\text{conf}}, c)$ (Eq. 4);
- $p_{\text{rewrite}}$ — conflict-focused query rewriting (Eq. 5);
- draft answer generation (Eq. 8);
- LLM judge (§C.2).

```
Look at this image carefully and identify the main subject entity.
Provide the Wikipedia article title for this entity (e.g., 'Eiffel
Tower', 'Great Wall of China', 'Siberian Husky'). Be specific and use
the exact Wikipedia page title format. Output ONLY the entity name in
<entity></entity> tags.
```

```
Step1 / Step2: Here is the question:{question}, Give your answer
and put the feature or reason for the answer related to the image
in <reason></reason>. Here is the question: {question}, Here is the
selected section: {sec[:800]}, Give your answer and put the feature
or reason for the answer related to the image in <reason></reason>.
Step3 : Here is the question:{question}, Below are the reasons
supporting the answer derived from the retrieved information:
{reasons} Identify the key features that distinguish the
aforementioned categories and extract the features that need attention
in <think></think>, without detailed illustration or answer.
Step4: Question: {question}, Conflict analysis:{R_vis}, Rate conflict
severity (0=no conflict, 10=completely contradictory). Output only a
number in <score></ score>.
```

```
You are a knowledge retrieval expert helping resolve factual conflicts
in visual QA.
Original question: {question}
Previous search queries (DO NOT repeat these): {history_str}
{entity_hint} Current conflict analysis: {R_vis}
Task: Generate a NEW, specific search query to find evidence that
resolves the conflict. Requirements: 1. Target the specific
disputed fact in the conflict 2. Must be different from all previous
queries 3. Be concise and keyword-focused (suitable for Wikipedia
search) 4. Focus on the entity/attribute that is in dispute
Output ONLY the new query wrapped in <query>...</query>.
```

```
Context: {final_text}
Question: {question} Short answer (1-5 words):
```

# C Evaluation Protocol

## C.1 LLM Judge

Following prior InfoSeek studies, every prediction is scored by an LLM judge (**DeepSeek-V4-Pro**, temperature 0) that decides whether the prediction is semantically correct with respect to the reference answer, rather than requiring exact string matching. The judge model, prompt, and decoding configuration are identical for every method — including the re-run CC-VQA baseline — and were never changed during evaluation.

| Metric | Value |
|---|---|
| Initial top-1 answer hit rate | 12.8% |
| Refined top-1 answer hit rate | 15.3% |
| Improvement | +2.5% |
| Newly promoted correct passages | 8.2% |

Table 5: Evidence quality before vs. after refinement (46,994 triggered samples).

Table 6: Comparison of aggregation strategies on InfoSeek 10K subset.

| Aggregation Strategy | Accuracy (%) |
|---|---|
| Geometric mean | 45.32 |
| Additive aggregation | **46.42** |

### C.2 JUDGE PROMPT

```
Judge if the prediction correctly answers the question with the same
meaning as ANY reference answer.
Question: {question} All reference answers (any one being correct):
{references} Prediction: {prediction}
Rules: - If the prediction contains any reference answer as a core
answer, it is YES. - "1830" / "in 1830" / "the year 1830" / "it was
built in 1830" -> YES (same answer 1830) - "Endangered" / "EN" /
"Endangered species" -> YES (same meaning) - "Pakistan" / "Islamic
Republic of Pakistan" / "PAK" -> YES (same entity) - "South Africa" /
"The vehicle belongs to South Africa" -> YES (contains correct answer)
- A longer sentence that contains the correct answer is still YES. -
"32.5" vs "14 to 33" -> NO (different specific value) - "Urca" vs
"Rio de Janeiro" -> NO (not specific enough, different place) - If
prediction is empty, refuses to answer, or gives irrelevant info -> NO
Answer ONLY "1" (YES) or "0" (NO).
```

## D EXTENDED RESULTS

### D.1 EVIDENCE REFINEMENT ANALYSIS

To directly evaluate the proposed evidence refinement mechanism, we analyze evidence quality before and after refinement on the 46,994 samples where conflict-aware rewriting was triggered (65.9% of all samples). As shown in Table 5, after conflict-aware refinement, the top-1 answer hit rate improves by 2.5 percentage points.Additionally, in 8.2% of triggered samples, conflict-guided reranking promotes correct passages from the original top-K candidate pool that were absent from the initial top-$N$ evidence set.

### D.2 EFFECTIVENESS OF EVIDENCE RELIABILITY MODELING

We further compare additive aggregation (used in CAER) with geometric aggregation $(r_q r_e r_v r_c)^{1/4}$: As shown in Table 6, additive aggregation outperforms geometric by 1.1 percentage points. Under geometric aggregation, a single near-zero signal drives the product toward zero and discards the information carried by the remaining signals; additive aggregation preserves each signal's independent contribution and is more robust in practice.

### D.3 RESULTS ON OK-VQA

We additionally evaluate CAER on OK-VQAMarino et al. (2019) (Table 7), scored with the same judge protocol of Section C.1. For a controlled comparison we re-ran the official CC-VQA code under our protocol: it scores 78.1 ($-0.7$ vs. the 78.8 reported under its original evaluation), so all gaps below are measured against the reproduced number under identical scoring. CAER reaches $\mathbf{81.4}$

while remaining training-free, on the same Qwen2.5-VL-7B backbone as CC-VQA — an exactly matched-backbone comparison.

Table 7: Performance on OK-VQA. Gen.FT = generator fine-tuned. †: reproduced with the official code under our protocol.

| **Method** | **Model** | **Gen.FT** | **Accuracy** |
|---|---|---|---|
| Qwen2.5-VL-7B | - | - | 72.4 |
| Wiki-PRF-7B | Qwen2.5-VL-7B | ✓ | 77.8 |
| MMKB-RAG | LLaMA-3.1-8B | ✗ | 65.4 |
| KU-RAG | LLaVA-Next-7B | ✗ | 73.1 |
| CC-VQA | Qwen2.5-VL-7B | ✗ | 78.8 |
| CC-VQA† | Qwen2.5-VL-7B | ✗ | 78.1 |
| **CAER (Ours)** | Qwen2.5-VL-7B | ✗ | **81.4** |