Identifying and Resolving Pitfalls of Knowledge-Based VQA Benchmarks: Auditing, Repairing, and Augmenting

Qian Ma[1], S M Rayeed[1], Charles V. Stewart[1], Qiong Wu[2], and Yao Ma[1]

[1] Rensselaer Polytechnic Institute, Troy NY 12047, USA
[2] AT&T Chief Data Office, Bedminster NJ 07921, USA
{maq5,rayees,stewart,may13}@rpi.edu, qw6547@att.com

Abstract. Knowledge-Based Visual Question Answering (KB-VQA) aims to evaluate whether Visual Language Models (VLMs) can retrieve, ground, and reason over external structured knowledge beyond visual evidence. In practice, answer accuracy is widely adopted as the primary evaluation metric, implicitly treating correctness as a proxy for knowledge-grounded reasoning. However, for existing KB-VQA benchmarks, this proxy relies on critical assumptions that are often overlooked and rendered unreliable by benchmark issues: annotated answer must be derivable from the associated knowledge base, question must be well-posed with sufficient constraints, and visual setting must meaningfully require grounded disambiguation. In this work, we show that these assumptions are systematically violated in existing KB-VQA benchmarks. Our audit reveals substantial instances with missing or contradicted answers and underspecified questions that render accuracy a misleading metric. Furthermore, we find that existing datasets rely on visually trivial, single-entity scenes that bypass the need for sophisticated visual-to-knowledge mapping. We demonstrate that even with controlled architectures, these flaws lead to distorted model rankings and overestimations of reasoning capabilities. To address this, we introduce (1) a principled audit-and-repair protocol that restores answer derivability and question clarity, and (2) a controlled multi-entity augmentation protocol that introduces visual ambiguity to challenge initial retrieval and grounded reasoning. Re-evaluation under corrected and augmented settings yields markedly different performance trends. Our findings call for rethinking evaluation protocols and designing more interaction-aware KB-VQA benchmarks that prioritize verifiable reasoning over simple matching. [1]

Keywords: KB-VQA · VLM · RAG

[1] The datasets and code are available in https://github.com/VAN-QIAN/ECCV26-ARA. Work was initiated when Qian was an intern at AT&T CDO. Qiong Wu and Yao Ma are co-corresponding authors.

1 Introduction

Recent Vision-Language Models (VLMs) [1,3,29,33] have demonstrated strong performance on a variety of visual reasoning tasks, especially Visual Question Answering (VQA) settings that require aligning image content with language understanding [2,11,15,16]. Despite such effectiveness, they are still struggling to address tasks where answer is beyond the visual part like Knowledge-Based Visual Question Answering (KB-VQA). KB-VQA [6,22] is designed to evaluate whether a model can answer image-grounded questions correctly by retrieving and reasoning over an external, controlled knowledge base, rather than relying only on parametric memory [6, 9], where accuracy is used to assess such knowledge-grounded capability.

Emerging efforts are focusing on achieving better answer accuracy by developing re-ranker modules to select the most relevant and informative evidence segment for the answer generation [20,31], improving the model's inherent capability to use all retrieved knowledge [7,8,13,30] or empowering VLMs to invoke external tools to retrieve more relevant evidence [14]. Significantly increasing efforts and resources are being invested to achieve better accuracy score.

However, in representative KB-VQA benchmarks such as InfoSeek and E-VQA [6,22], we observe recurring dataset issues (Section 3) that can make answer accuracy an unreliable indicator of the intended knowledge-grounded reasoning capability. **1. Answer-evidence misalignment.** A non-trivial portion of questions have annotated answers that are not supported by the provided knowledge resource (*e.g.* missing or contradicted). As shown in Figure 1, when answering the sea level of the memorial park, the desired answer is '10 foot' while it's missing in the benchmark-provided knowledge base. **2. Underspecified questions and answer scope.** Many questions do not specify the intended answer granularity, so multiple answers can be reasonable. For example, for question "When is the mating period of this bird?", the evidence may mention a season (Spring), specific months (March-April), or a life-stage (around 1-year old). When several evidence-supported answers are plausible, accuracy becomes sensitive to the single particular annotated answer, and can therefore understate a model's knowledge-grounded capability. **3. Visually simplified scenes vs. real-world multi-modal queries.** Existing KB-VQA benchmarks [6,22] present visually simplified scenes dominated by a single salient entity, so questions can often be answered using coarse global cues and retrieval may succeed without explicit grounding to the intended target [23,25]. In realistic scenarios, however, users often issue more complex multi-modal queries that require text-image interaction *i.e.* the query text specifies which entity is relevant (*e.g.*, via spatial relations, relational descriptions, *etc.*), and the system should localize candidate entities and resolve the intended entity before KB retrieval and reasoning. For example, a query may refer to "the fish on the right" where multiple plausible entities coexist and global image similarity alone is insufficient.

These issues matter because they decouple benchmark scores from the capability KB-VQA aims to measure. Under such conditions, high answer accuracy

can be driven by annotation artifacts, dataset biases, or shortcut strategies, rather than faithful knowledge-grounded reasoning.

To address these limitations, we propose a unified fixing protocol and augmentation framework. The fixing protocol enforces answer derivability and question clarity via evidence verification, answer-source consistency checks, and targeted question repair (Section 4). Building on the repaired benchmark, we further introduce a controlled augmentation pipeline that injects visual distractors while preserving core QA semantics, creating more challenging multi-entity settings for grounded retrieval and reasoning (Section 5).

Re-evaluation with representative KB-VQA baselines [7,14,20,30,31] reveals that correcting dataset validity alone can materially change measured performance and method ranking, while augmentation causes substantial drops in retrieval recall and end-to-end QA accuracy (Section 6). These findings suggest that current KB-VQA evaluation can be overly optimistic, and that robust progress requires both annotation-level validity and explicit grounding difficulty.

2 Related Works

2.1 Existing KB-VQA Datasets

The transition from general VQA [2,11] to KB-VQA [6,22] marks a shift from internal visual perception to the integration of open-world knowledge. While early benchmarks like OK-VQA [21] and FVQA [28] focused on general commonsense reasoning, modern datasets have scaled toward massive, external knowledge sources to simulate more complex grounding. Concurrent works such as **InfoSeek** [6] and **E-VQA** [22] are two primary examples of this large-scale shift. InfoSeek introduces over 1.3 million samples to simulate real-world intent; however, it suffers from a paradigm flaw where the knowledge base often lacks the specific facts used during human annotation, leading to misalignment and evaluation distortion. E-VQA attempts to address this by providing a controlled corpus of 2 million Wikipedia pages with section-level evidence. Despite these improvements, both datasets rely heavily on template-based question generation, which often lacks the linguistic variety found in natural queries. Beyond these two datasets, there is a following work SK-VQA [24] that claims context comprehension capability is important to achieve KB-VQA task. Therefore, they constructed a synthetic dataset to generate 'pseudo' wiki articles to explicitly train the model to seek information among it. Beyond task formulation, prior KB-VQA benchmarks [6,22] also differ in how their knowledge sources, annotations, and question templates are constructed. These design choices can introduce recurring issues such as mismatches between annotated answers and retrievable evidence, underspecified templates that admit multiple plausible answers, and visually simplified scenes that weaken the need for grounded disambiguation. Our work builds on these benchmarks and provides a structured characterization of such issues and their implications for evaluation.

2.2 MM-RAG Solutions for KB-VQA

Since the answers of KB-VQA questions are beyond the visual part and external knowledge is required, most emerging solutions for KB-VQA tasks can be categorized into Multi-modal Retrieval-Augmented Generation (MM-RAG) frameworks to effectively acquire the external knowledge base according to the multi-modal queries. These solutions generally focus on how to selectively utilize retrieved content to respond to queries and can be categorized into two streams.

1. Re-ranking based methods mainly focus on picking the most relevant and informative part of all initial-retrieved information *e.g.* the section corresponding to target question of all retrieved Wikipedia articles. **IBA** [20] addresses this by explicitly decoupling the workflow to ground the target entity by doing identification before reasoning to answer. It leverages the zero-shot recognition capabilities of MLLMs to identify candidate entities before employing a lightweight text-re-ranker for precise evidence selection. This modular workflow re-design approach contrasts with framework **EchoSight** [31], which trains a customized multi-modal re-ranker based on Q-Former [17].

2. Aggregation/Filtering-based methods focusing on digesting all retrieved information to implicitly provide the evidence for downstream answer generation instead of explicitly selecting the evidence. **ReflectiVA** [7] introduces specialized self-reflective tokens that allow the re-trained MLLM to autonomously determine the necessity of each retrieved section. All sections considered as relevant will be aggregated to support the answer generation. **CoMeM** [30] proposed to condense all retrieved articles into the embeddings and concatenate with query embeddings to support the answer generation as model memory. In **Wiki-PRF** [14], reinforcement learning is introduced to empower MLLMs to conduct multiple retrievals with different tools and filtering all retrieved results to support the final answer generation.

3 Issues in Existing KB-VQA Benchmarks

Answer accuracy is widely reported on KB-VQA benchmarks, but it can be unreliable when benchmark instances violate basic validity requirements. Across E-VQA [22] and InfoSeek [6], we observe three recurring dataset issues that may directly distort evaluation: (i) **answer-evidence misalignment** that creates unavoidable false negatives, (ii) **underspecified questions** where multiple answers are defensible and scores depend on the single desired annotated answer, and (iii) **visually simplified scenes** where shortcut retrieval can succeed without grounded disambiguation. Together, these issues can decouple benchmark scores from the knowledge-grounded retrieval and reasoning capability KB-VQA is intended to measure. We use three assumptions, *(A) Answer Derivability, (B) Question Well-Posedness,* and *(C) Grounded Disambiguation Requirement* as an organizing lens, and document representative violations under this framework.

3.1 Answer-Evidence Misalignment

Assumption (A) requires that each annotated answer be derivable from the associated KB evidence. We observe two recurring violations: *(a) Unsupported answers*, where the answer is absent from the KB (*e.g.* InfoSeek QID 23994 annotates "10 foot" for sea level of Wright Brothers National Memorial while corresponding evidence article doesn't contain any relevant information), and *(b) Contradictory answers*, where annotation conflicts with evidence (*e.g.* InfoSeek QID 5411 annotates "1,302" kilogram as Mass of McLaren car, but evidence states "It weighs 1,301 kg" such a single value won't match the desired range).

When derivability fails, evaluation produces unavoidable false negatives under the benchmark-provided KB that even perfect retrieval and correct reasoning can be scored as wrong because the target annotation is not grounded in the retrievable evidence (Figure 1). Using an initial verification pass with Qwen3-30B-A3B [32] to scan each target entity page and judge support and derivability, we find unsupported annotations in about 1% of E-VQA and 22% of InfoSeek. This implies that a non-trivial fraction of InfoSeek is unanswerable under its own provided KB, capping accuracy by construction. This failure is largely structural in InfoSeek, where QA pairs come from WikiData [27] knowledge graph triplets but evaluation uses Wikipedia text, creating a systematic cross-source mismatch.

3.2 Underspecified Questions

Assumption (B) requires each question to be sufficiently constrained so that the annotated answer is unique under the given KB. We identify three frequent ambiguity patterns:
(1) Missing attribute constraint e.g. For the question "How big can this plant become?" without specifying height or diameter.
(2) Missing temporal scope e.g. "When is the mating period?" without phase or time granularity, then the answer could be a season(spring) or month(September) during the year or during the species' life-cycle(such as '1-year old').
(3) Missing spatial reference or granularity (*e.g.* For the same question "In which country or region does this animal live?", the desired habitat answers could be region-level(North America) or country-level(United States)).

Using these tags, our audit shows that 59% of E-VQA and 47% of InfoSeek questions are ambiguous: E-VQA has 21.5% attribute, 27.5% spatial, and 10% temporal omissions; InfoSeek shows a similar pattern (17.3%, 30.3%, 2%). Under such under-specification, models may produce semantically valid alternatives but still be penalized as wrong. Therefore, accuracy becomes sensitive to annotation preference rather than reasoning correctness. These ambiguities mainly stem from template-based generation [22] and automatic KG-to-question mapping [6], where key qualifiers are dropped.

3.3 Single-Entity Shortcut and Missing Grounded Disambiguation

Assumption (C) requires that success depends on grounded disambiguation to respond to complex multi-modal user queries under realistic scenarios. Ideally,

KB-VQA involves four stages: entity localization, text-image grounding, entity disambiguation, and KB retrieval/reasoning. In realistic queries (*e.g.* "the fish on the left"), the model should first identify which entity is being referenced. However, existing KB-VQA benchmarks usually contain one dominant entity per image and evaluate against a known target identifier. Because the target is effectively fixed per instance, global retrieval can succeed without verifying which entity the question refers to. With this design, global image embeddings can often retrieve the target directly, collapsing localization, grounding and disambiguation into a shortcut. As a result, high answer accuracy does not necessarily imply robust grounding ability. The metric can overestimate capability because shortcut retrieval is sufficient in many samples. This issue is driven by entity-centric curation, target-anchored evaluation protocols, and reliance on global embedding retrieval [6, 22]. To restore this missing requirement, Section 5 introduces controlled multi-entity augmentation so that localization, grounding, and disambiguation become necessary for success.

4 Proposed Fixing Protocol

To address recurring benchmark issues such as **answer-evidence mismatch** and **underspecified** questions (Section 3), we propose a dataset-level audit-and-repair protocol applicable to KB-VQA benchmarks. It explicitly enforces the two assumptions identified in Section 3: *Assumption A (Answer Derivability):* the annotated answer must be supported by and derivable from the external knowledge base; *Assumption B (Question Well-Posedness):* the question must provide sufficient constraints to uniquely determine the annotated answer. Given a KB-VQA instance (image, QA pair, target entity, and KB evidence snapshot), the pipeline outputs either a repaired instance or a filtered-out instance. It consists of four cascaded stages: evidence verification, answer-derivability auditing, question-constraint repair, and final consistency validation.

4.1 General Four-Stage Protocol

Step 1: Evidence Verification. We first verify whether the evidence referenced by each QA pair exists in the KB snapshot and whether it contains support for the annotated answer under the target entity context. Instances are categorized as **Supported & Matched**, **Supported but Mismatched**, and **Unsupported**. Since InfoSeek constructs QA pairs from Wikidata while using Wikipedia articles as the evaluation KB, we scan the target-entity page section by section and apply two independent verifiers (Qwen3-30B-A3B [32] and DeepSeek-v3.2 [18]) to mitigate erroneous filtering due to cross-source mismatch. Only instances that are consistently classified as **Unsupported** by both verifiers are removed (Figure 1). The verification is constrained to checking whether the target-entity evidence supports, contradicts, or lacks the annotated answer, rather than making an open-ended judgment. The evidence context is localized to sections, with mean/95th-percentile lengths of 868/2360 tokens for E-VQA

and 877/2545 tokens for InfoSeek. For E-VQA, where evidence segments are provided during dataset construction, we directly examine the referenced segment and revisit cases with answer–evidence mismatch.

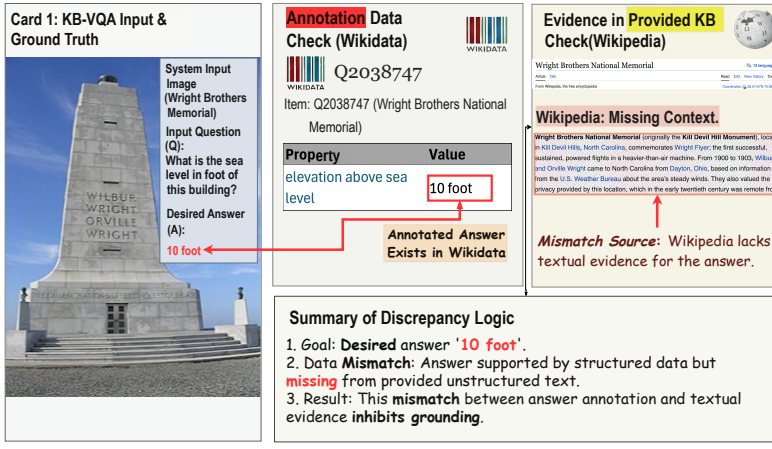

Fig. 1: Qualitative example from InfoSeek (QID:149). InfoSeek [6] selects answers from Wikidata [27] triples and converts them into QA pairs, while the evaluation KB consists of Wikipedia articles. This cross-source construction can yield cases where the provided KB does not support the annotated answer, confounding score interpretation even under perfect retrieval.

Step 2: Answer-Derivability Auditing and Calibration. Given verified evidence and the original question, we test whether the annotated answer is logically derivable. For any instances tagged as **Supported but Mismatched**, if the answer contradicts evidence, we apply evidence-grounded answer correction. DeepSeek-v3.2 is not used as an answer source. It only rewrites a mismatched annotation to the value explicitly supported by verified evidence. If no evidence-supported value can be derived, the instance is removed. For InfoSeek, this step frequently corrects answer-side mismatch introduced by cross-source construction. For E-VQA [22], answer-evidence contradictions are much less frequent, consistent with its construction [4] where the answer is selected from the provided evidence article.

Step 3: Question-Constraint Repair. We identify whether the question is sufficiently constrained for unique answering. We repair ambiguous questions with minimal edits using three tags: missing attribute constraints, missing temporal scope, and missing spatial reference. Edits preserve the original intent and evidence dependency. For InfoSeek, the dropped qualifiers from KG-to-question mapping are restored by adding missing facets. For E-VQA, the original template-generated questions to fit super-category are revised mainly when templates omit disambiguating qualifiers to the fine-grained entity.

Step 4: Leakage and Global Consistency Validation. After repair, we run a final pass to ensure that: (i) the question does not leak the answer, (ii) the answer remains evidence-supported, and (iii) the revised question yields a unique answer. Any leakage-inducing edits are rolled back and rewritten conservatively.

4.2 Fixing Outcomes and Human Evaluation

Applying the same four-stage protocol to both datasets yields different failure profiles but a unified corrected setting. For InfoSeek, we retain 58,285 out of 71,335 instances (81.7%) after filtering and repair. We construct an entity-unique subset of 1,924 questions(consisting 1604 String, 223 Numerical and 97 Time questions) for controlled evaluation *i.e.* for instances targeting on the same target entity with same query text but different query images, we only keep one. For E-VQA, answer-side conflicts are rarer and most edits happen in question constraints. Hence the fixed evaluation split remains 4,750 questions. For human evaluation, we expand the review to 10% of the fixed evaluation splits, covering 475 E-VQA and 200 InfoSeek instances. Following existing work [24], two PhD-student-level annotators answer each fixed query using only the oracle evidence of the target entity. Human accuracy remains high: 92.9 ± 1.1 for fixed E-VQA and 91.5 ± 2.1 for fixed InfoSeek. More qualitative examples of fixing outcomes and human-evaluation cases are provided in Appendix A and Appendix F. These repaired datasets are used as the *fixed* versions and original datasets with the same index/identifier are used as the *unfixed* in Section 6.

5 Proposed Augmentation Protocol

As discussed in Section 3, a gap remains between current KB-VQA benchmark settings and real-world multi-modal queries, where users often must specify the target entity through text alone under multi-entity ambiguity (*e.g.*, "the fish on the right"). In existing benchmarks, however, images are frequently dominated by a single salient entity, so retrieval can succeed via coarse global image similarity overlooking *(C) Grounded Disambiguation Requirement*.

5.1 General augmentation protocols

To close the evaluation gap in Section 3, we augment each instance by adding a single distractor entity alongside the original target (anchor). We preserve the annotated answer by construction and keep the external knowledge source unchanged, so performance changes reflect increased visual ambiguity rather than knowledge or annotation shifts. Each augmented image contains exactly one anchor and one distractor, where the distractor is drawn from either the same semantic category (intra-category) or a different one (inter-category), forcing retrieval to rely on grounded disambiguation rather than coarse global similarity.
Intra-Category Distraction We spatially concatenate the anchor image with a single distractor from the same semantic category, placed in a different region

(*e.g.*, left vs. right). We minimally revise the question to reference the anchor region or a distinguishing attribute (*e.g.*, "the fish on the left"), while keeping the annotated answer unchanged. This setting isolates fine-grained intra-category ambiguity and tests whether models ground textual constraints in localized visual evidence rather than relying on coarse global embeddings.

Inter-Category Distraction We pair the anchor entity with a single distractor from a different semantic category (*e.g.*, a landmark with an animal, or a plant with a man-made object). Although visually distinct, the distractor can interrupting global image representations, requiring models to integrate the question with localized evidence to infer the intended referent. The question and answer remain unchanged, allowing us to test whether retrieval is overly sensitive to irrelevant visual content instead of grounding target entity under textual constraints.

5.2 Augmentation Outcomes and Human Evaluation

To enable direct comparisons, we exclude questions that would remain ambiguous after intra-category augmentation (*e.g.*, those referring to "the object") from *both* splits. As shown in Figure 2, intra-category augmentation adds a minimal cue (typically spatial) to specify the anchor entity, whereas inter-category augmentation keeps the question unchanged. With answers preserved by construction, performance changes reflect increased grounded disambiguation demands rather than knowledge or annotation shifts. We generate 1,604 (InfoSeek) and 3,871 (E-VQA) augmented variants, with additional examples in Appendix B. For quality assurance, we sample 100 intra- and 100 inter-category instances per dataset. Two annotators answer each augmented query given the anchor evidence and optionally report evident flaws [24]. Human accuracy is $87.5 \pm 2.1\%$/ $96.0 \pm 1.4\%$ (E-VQA intra/inter) and $86.0 \pm 1.4\%$/ $89.5 \pm 3.5\%$ (InfoSeek intra/inter). More details are in Appendix G.

6 Experiments

6.1 Experimental Setup

Datasets. We evaluate KB-VQA baselines on the following aligned dataset variants: (1) the original *unfixed* benchmark, (2) a *fixed* version that enforces Assumption A/B via the pipeline in Section 4, and (3) the *Intra* and *Inter* augmented versions that enforces the grounded-disambiguation requirement via the pipeline in Section 5. Overall, these controlled comparisons tie each analysis directly to the assumptions in Section 3 and validate whether fixing and augmentation restore the intended evaluation conditions.

External Knowledge Base. E-VQA provides a controlled knowledge base of 2 million Wikipedia articles with associated images. Following prior work [31], we focus on the single-hop setting. For InfoSeek, following existing baselines [7, 14, 26], we adopt the 100K-article knowledge base released by Yan and Xie [31] and follow the same initial retrieval settings using EVA-CLIP-8B [25].

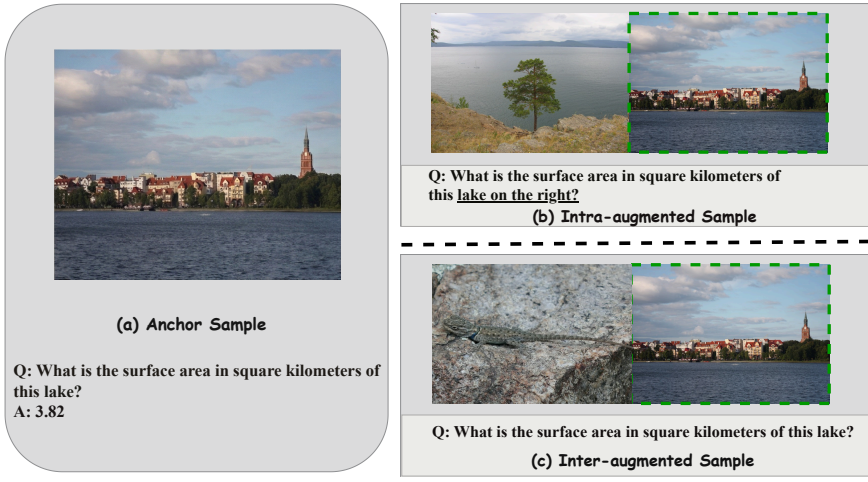

Fig. 2: Qualitative example of our augmentation protocols (Section 5) on *Ełk Lake*. All augmented variants keep the anchor answer in (a). (b) Intra-augmentation adds a minimal spatial cue ("on the right") to preserve the original intent while inserting an additional lake image (*Lake Turgoyak*) on the left. (c) Inter-augmentation inserts a visually distinct distractor (*Sceloporus jarrovii*) while keeping the question unchanged.

Baselines. We benchmark five representative KB-VQA methods, grouped by how retrieved evidence is utilized. *Re-ranking-based methods* (IBA [20], EchoSight [31]) explicitly re-rank retrieved evidence and generate the final answer conditioned on the top-ranked section, using either a tailored workflow paradigm or a trained multimodal re-ranker. *Aggregation/Filtering-based methods* (ReflectiVA [7], CoMeM [30], Wiki-PRF [14]) instead aggregate evidence across multiple sections or retrieval rounds and filter irrelevant content before answer generation. ReflectiVA assigns section-level special relevant tokens and aggregates the selected evidence. CoMeM encodes retrieved articles into memory representations that are consumed by an MLLM. Wiki-PRF performs iterative retrieval with auxiliary tools (*e.g.* grounding and captioning) and filters retrieved content across rounds to serve as evidence for answer generation.

Evaluation Metrics. We report both end-to-end QA accuracy and retrieval-related metrics. For answer accuracy, we evaluate open-ended outputs using BEM [34] on E-VQA. For InfoSeek, following prior practice [7, 14, 31], we use VQA accuracy [11, 14, 21]. For retrieval, we report Recall@K, indicating whether the correct target-entity article appears among the top-K retrieved results. For methods that include an explicit re-ranking stage [31] or multi-round retrieval [14], we additionally report *post-retrieval* performance, measuring whether the final evidence used for answer generation (*e.g.* the top-1 re-ranked section or the aggregated evidence pool) contains the correct target entity.

Implementation Details. For all baselines [7, 14, 30, 31], we use officially released checkpoints and prompts with unified KB and retrieval settings. For

IBA [20] component selection, we use Qwen2.5-VL-7B-Instruct [3] for identification, bge-reranker-v2-m3 [5] for re-ranking and LLama-3.1-8B-Instruct [12] for answer generation. Hyper-parameters follow the default configurations provided by the authors and are summarized in Appendix H. For initial retrieval, we use a frozen EVA-CLIP-8B [25] encoder to extract image features, following existing works [7,14,31]. All image features are indexed and retrieved with cosine similarity using Faiss-GPU library [10], consistent with prior practices [7,14,31].

6.2 Results on fixed datasets

We first evaluate all methods on both the original and fixed versions of InfoSeek and E-VQA. Importantly, we ensure the evaluated samples from both original and fixed datasets preserve exactly the same sample IDs. We only correct flawed question formulations and annotated answers. Therefore, any performance difference reflects violations of *Answer Derivability* and *Question Well-Posedness* (Section 3) rather than model changes.

Table 1: Performance comparison on InfoSeek and E-VQA before and after dataset fixing. For InfoSeek, most of methods performance improved, especially on Time and Numerical. **Bold** for best performance and underline to the runner-up

Methods	Backbone	InfoSeek								E-VQA	
		Unfixed				Fixed				Unfixed	Fixed
		All	Time	Num	String	All	Time	Num	String	All	All
Wiki-PRF [14]	Qwen	**44.9**	35.1	**45.7**	**45.3**	**43.6**	43.3	52.9	**42.3**	31.9	33.1
CoMEM [30]	Qwen	24.3	16.5	19.3	25.5	23.5	17.5	21.1	24.3	14.1	13.4
ReflectiVA [7]	Llama	37.3	37.1	43.9	36.4	38.1	34.0	58.7	35.5	36.8	36.6
IBA [20]	Llama	34.5	**38.1**	43.0	33.0	42.4	**47.4**	**61.9**	39.4	**41.9**	**42.7**
EchoSight [31]	Llama	30.9	29.9	38.1	30.0	37.2	32.0	51.1	35.6	40.4	40.9
LLaVA-v1.5 [19]	Llama	5.8	0.0	7.2	6.0	6.6	1.0	9.0	6.6	12.9	12.7
Qwen2.5-VL [3]	Qwen	21.4	12.4	14.3	22.9	28.3	9.3	17.0	31.0	21.9	21.9

By repairing QA annotations to enforce answer–evidence alignment and resolving underspecified questions, fixing yields substantial score changes across methods on InfoSeek. Because the fixed and unfixed evaluations are aligned on the same sample IDs, these shifts reflect the impact of removing annotation- and template-induced confounders rather than changes in evaluation coverage. Some methods improve markedly (*e.g.* IBA 34.5 $\rightarrow$ 42.4, EchoSight 30.9 $\rightarrow$ 37.2), with larger gains on the stricter Time and Num subsets (*e.g.* Wiki-PRF 35.1 $\rightarrow$ 43.3, 45.7 $\rightarrow$ 52.9). Because images, retrieval settings, checkpoints, and metrics are held constant, these shifts isolate the impact of benchmark validity (Assumption A/B) rather than changes in model design. Crucially, we also observe ranking reversals after fixing, indicating that unfixed splits can distort comparative conclusions about which system components matter most.

Key observations from fixing results. *(i) Relative gaps shrink after fixing.* On unfixed InfoSeek, aggregation/filtering pipelines appear much stronger than re-ranking based methods like IBA (*e.g.* Wiki-PRF vs. IBA: +10.4; ReflectiVA vs. IBA: +2.8). After fixing, these gaps shrink to +1.2 and −4.3, respectively.

(ii) Method rankings can reverse. On unfixed InfoSeek, ReflectiVA outperforms IBA, but the ordering reverses after fixing (IBA 42.4 vs. ReflectiVA 38.1). On the Time subset, explicit evidence selection becomes more favorable after fixing (IBA 47.4 and Wiki-PRF 43.3 vs. ReflectiVA 34.0). These reversals matter in practice because they can change system-design conclusions. For example, after the fixing, community may draw a conclusion that selecting the most relevant and informative evidence section is more effective than relying on model's inherent capability to assess the relevance of each evidence section to query. Otherwise, if the community sticks to the conclusion draws from the unfixed data, a lot of efforts could be invested to expensive training but the gain may not aligned in a corresponding level. Similarly, the advantage of Wiki-PRF over IBA is large on unfixed InfoSeek (44.9 vs. 34.5, gap 10.4), but becomes much smaller after fixing (43.6 vs. 42.4, gap 1.2). In this case, community may reconsider the trade-offs between developing new models and reconsidering the workflow orchestration. Overall, violations of answer derivability and question well-posedness can distort comparative conclusions and misallocate optimization effort.

Due to space limits, we defer detailed fine-grained and qualitative analyses to Appendix D to show how our fixing improves the performance by correcting the desired annotation answer and improving the question to guide the desired answer. On the shared grounded InfoSeek subset where EchoSight's top-1 reranked section belongs to the ground-truth entity (664 samples), question-only fixes improve QA accuracy from 60.1 to 75.1, joint question-answer fixes improve from 34.0 to 45.2, and answer-only fixes remain nearly unchanged. Detailed breakdowns are in Appendix D. Post-retrieval results also provide supporting evidence for re-ranking methods that improving the question contributes to better re-ranking. On InfoSeek, post-retrieval performance increases after fixing for EchoSight (46.8 → 48.9) and IBA (46.7 → 47.8). This suggests that part of the QA improvement is associated with better evidence selection after fixing, while question and answer repairs likely contribute jointly. Table 6 shows that InfoSeek has high modification rates for both questions and answers, whereas E-VQA mainly changes questions. This aligns with the observed pattern: E-VQA shifts are modest, while InfoSeek shows larger magnitude changes in Table 1.

Retrieval performance affects cross-dataset rankings. We further report retrieval recall on the fixed full datasets and the anchor/augmented variants to analyze retrieval difficulty and to contextualize augmentation effects. Table 2 shows that InfoSeek has much higher initial R@1 than E-VQA (43.5% vs 13.4% on the full fixed set). Correspondingly, Table 3 shows that Wiki-PRF remains competitive on InfoSeek (48.6%) but drops sharply on E-VQA (18.2%), while EchoSight and IBA are more stable. This pattern indicates that ranking differences across datasets are driven largely by retrieval difficulty rather than by

Table 2: Initial retrieval recall. Since initial retrieval are following the same image-to-image protocol of all baselines, the fixed and unfixed version share the same retrieved information. For Intra- and Inter-augmented version, augmented images are used.

Version	E-VQA					InfoSeek				
	R@1	R@3	R@5	R@10	R@20	R@1	R@3	R@5	R@10	R@20
Full	13.4	26.1	31.9	41.8	48.9	43.5	60.1	65.8	72.1	75.8
Anchor	12.8	26.3	32.6	43.3	50.7	43.1	59.9	66.1	72.7	76.4
Intra-Aug	3.5	7.5	9.8	13.6	17.0	14.7	24.4	34.4	36.5	41.3
Inter-Aug	2.9	6.7	8.6	12.3	15.0	20.4	29.5	31.9	34.2	34.4

downstream reasoning, which further strengthen the need to challenge the retrieval as we discussed in Section 5.

Table 3: Post-retrieval performance using retrieved **top 20** articles. on all dataset variants, including unfixed, fixed datasets, the subset of anchor samples and the augmented versions. Note for ReflectiVA [7] and CoMEM [30], they aggregate the top-5 and top-10 articles *i.e.* align with initial retrieval recall@5 and recall@10 in Table 2.

Method	E-VQA					InfoSeek				
	Unfixed	Fixed	Anchor	Intra	Inter	Unfixed	Fixed	Anchor	Intra	Inter
EchoSight	36.4	36.3	46.6	11.8	11.7	46.8	48.9	48.4	16.8	20.4
IBA	34.7	35.0	44.8	8.9	9.1	46.7	47.8	45.1	21.7	21.6
Wiki-PRF	18.0	18.2	17.9	4.9	4.7	48.2	48.6	48.3	18.1	17.6

6.3 Results on augmented datasets

To directly test the grounded-disambiguation requirement (Section 5), we evaluate on the augmented variants that introduce a single distractor while preserving the answer and KB. We report three representative methods (EchoSight, IBA, Wiki-PRF) to cover re-ranking and aggregation paradigms.

Assumption C (Grounded Disambiguation): augmentation removes the single-entity shortcut. Table 2 shows that adding a single distractor sharply reduces initial retrieval recall. On InfoSeek, R@1 drops from 43.5 (Full) to 14.7/20.4 (Intra/Inter); on E-VQA, it falls from 13.4 to 3.5/2.9. Since KB and questions are unchanged, this collapse is attributable to the added entity ambiguity, confirming the augmentation effectively enforces grounded disambiguation. To separate layout effects from semantic distractors, we additionally evaluate two layout-only controls: *Blank* replaces the distractor with an empty panel and *Double* duplicates the anchor image. Initial retrieval is layout-sensitive, but Blank/Double remain easier than Intra/Inter (E-VQA R@1: 6.3/9.9 vs. 3.5/2.9

and InfoSeek R@1: 26.2/36.9 vs. 14.7/20.4). Full initial- and post-retrieval results are reported in Appendix E, showing that semantic distractors introduce difficulty beyond layout shift.

Post-retrieval evidence selection does not recover the loss. Table 3 shows that post-retrieval recall also degrades sharply on augmented images (*e.g.* IBA InfoSeek 45.1 → 21.7/21.6, EchoSight E-VQA 46.6 → 11.8/11.7). This suggests that once initial grounding fails, later evidence aggregation cannot compensate, aligning with our diagnosis that grounded disambiguation is the critical bottleneck. Surprisingly, the Wiki-PRF didn't try to actively invoke its grounding tool. We defer more detailed analysis in Appendix E.

QA accuracy drops sharply, validating the lack of grounded disambiguation. Table 4 reports QA accuracy on the anchor subset and its two augmented variants. All methods suffer substantial performance degradation once ambiguity is introduced. For example, on InfoSeek, IBA drops from 40.1 (Anchor) to 21.4/21.6 (Intra/Inter), and EchoSight drops from 38.7 to 15.9/17.8. Even Wiki-PRF declines from 43.9 to 23.6/25.9. Similar trends hold on E-VQA.

Table 4: QA Performance on Intra-category(**Intra**) and Inter-Category(**Inter**) augmented datasets and the corresponding anchor samples. Following Table 1, we report the performance on E-VQA(**E-V**) [22] and InfoSeek(**IS**) [6] with breakdowns on Time(**T**), Numerical(**N**) and String(**S**) questions

Methods	Anchor					Intra					Inter				
	IS				E-V	IS				E-V	IS				E-V
	All	T	N	S		All	T	N	S		All	T	N	S	
IBA	40.1	45.8	25.9	41.7	38.4	21.4	16.9	17.5	22.2	19.8	21.6	28.9	14.8	22.1	16.6
Wiki-PRF	43.9	33.7	1.6	50.5	32.8	23.6	10.8	0	27.7	23.1	25.9	22.9	0	29.9	22.5
EchoSight	38.7	30.1	22.8	41.4	42	15.9	8.4	7.9	17.5	19.3	17.8	14.5	9.5	19.1	18.5

These results indicate that, even after repairing QA validity (Section 4), existing KB-VQA evaluations remain optimistic since their visual setups allow retrieval pipelines to exploit single-entity shortcuts. Our controlled augmentation thus exposes a remaining bottleneck: **how to reliably ground the query to the correct entity and then select truly relevant evidence under clutter**, which suggests that future KB-VQA benchmarks should incorporate richer multi-entity interactions and stronger grounding requirements.

7 Conclusion

We show that current KB-VQA benchmarks can exhibit recurring dataset issues that complicate evaluation by answer accuracy alone. Across InfoSeek and E-VQA, we identify three prominent confounders: (i) answer–evidence misalignment where annotated answers are not derivable from the benchmark-provided

knowledge snapshot, (ii) underspecified questions that admit multiple plausible answers, and (iii) visually simplified settings that weaken the need for grounded disambiguation. These issues can decouple scores from the knowledge-grounded retrieval and reasoning capability KB-VQA is intended to measure.

To address them, we propose two protocols. First, a fixing protocol that enforces answer derivability and question well-posedness via evidence-aware auditing and targeted repair. Second, a controlled augmentation strategy that introduces visual distractors while preserving the annotated answer, thereby increasing the need for grounded entity disambiguation under multi-entity conditions. Together, these protocols transform existing benchmarks into a more diagnostic testbed for retrieval-and-reasoning.

Experiments on the repaired and augmented splits reveal trends that differ from standard evaluation. After fixing, most methods improve substantially, and we observe that relative rankings can change, indicating that unfixed benchmarks may distort comparative conclusions and the perceived impact of system components. Under augmentation, performance drops consistently for both retrieval recall and end-to-end accuracy, highlighting that robust grounding and intent-aligned retrieval remain challenging under visual ambiguity.

More broadly, our findings suggest that KB-VQA evaluation should explicitly account for benchmark validity. We encourage the community to (i) prioritize evidence-derivable annotations and well-specified questions during dataset construction, (ii) report metrics that reflect evidence support and grounding robustness in addition to answer correctness, and (iii) develop retrieval and reasoning methods that remain reliable when multiple plausible entities and distractors are present. We hope our protocols and revised benchmarks facilitate more faithful, diagnostic, and reproducible KB-VQA evaluation.

Acknowledgements

This research was supported by the National Science Foundation (NSF) under grant numbers NSF2406647 and NSF2406648. It was also supported by the National Artificial Intelligence Research Resource (NAIRR) Pilot and the Delta advanced computing and data resource, which is supported by the National Science Foundation under award NSF-OAC-2005572. S. M. R. and C. V. S. are supported by the Imageomics Institute, funded by the US National Science Foundation's Harnessing the Data Revolution (HDR) program under Award No. 2118240 (Imageomics: A New Frontier of Biological Information Powered by Knowledge-Guided Machine Learning).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Changpinyo, S., Kukliansy, D., Szpektor, I., Chen, X., Ding, N., Soricut, R.: All you may need for vqa are image captions. In: Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: human language technologies. pp. 1947–1963 (2022)
5. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216 (2024)
6. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: Can pre-trained vision and language models answer visual information-seeking questions? arXiv preprint arXiv:2302.11713 (2023)
7. Cocchi, F., Moratelli, N., Cornia, M., Baraldi, L., Cucchiara, R.: 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 (CVPR). pp. 9199–9209 (June 2025)
8. Compagnoni, A., Morini, M., Sarto, S., Cocchi, F., Caffagni, D., Cornia, M., Baraldi, L., Cucchiara, R.: Reag: Reasoning-augmented generation for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11901–11911 (2026)
9. Deng, J., Wu, Z., Huo, H., Xu, G.: A comprehensive survey of knowledge-based vision question answering systems: The lifecycle of knowledge in visual reasoning task. arXiv preprint arXiv:2504.17547 (2025)
10. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. IEEE Transactions on Big Data (2025)
11. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
12. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
13. Hong, Y., Gu, J., Lou, Y., Fan, L., Yang, Q., Wang, Y., Ding, K., Wu, Y., Xiang, S., Ye, J.: 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)
14. Hong, Y., Gu, J., Yang, Q., Fan, L., Wu, Y., Wang, Y., Ding, K., Xiang, S., Ye, J.: Knowledge-based visual question answer with multimodal processing, retrieval and filtering. arXiv preprint arXiv:2510.14605 (2025)
15. Kim, B.S., Kim, J., Lee, D., Jang, B.: Visual question answering: A survey of methods, datasets, evaluation, and challenges. ACM Computing Surveys **57**(10), 1–35 (2025)
16. Kuang, J., Shen, Y., Xie, J., Luo, H., Xu, Z., Li, R., Li, Y., Cheng, X., Lin, X., Han, Y.: Natural language understanding and inference with mllm in visual question answering: A survey. ACM Computing Surveys **57**(8), 1–36 (2025)

17. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
18. Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al.: Deepseek-v3.2: Pushing the frontier of open large language models. arXiv preprint arXiv:2512.02556 (2025)
19. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
20. Ma, Q., Wu, Q., Zhou, Z., Ma, Y.: Ground then rank: Revisiting knowledge-based vqa with training-free entity identification (2026), https://arxiv.org/abs/2606.23881
21. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: 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)
22. Mensink, T., Uijlings, J., Castrejon, L., Goel, A., Cadar, F., Zhou, H., Sha, F., Araujo, A., Ferrari, V.: 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)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
24. Su, X., Luo, M., Pan, K.W., Chou, T.P., Lal, V., Howard, P.: SK-VQA: Synthetic knowledge generation at scale for training context-augmented multimodal LLMs. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=EVwMw2lVlw
25. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: Eva-clip-18b: Scaling clip to 18 billion parameters. arXiv preprint arXiv:2402.04252 (2024)
26. Tian, Y., Liu, F., Zhang, J., Hu, Y., Nie, L., et al.: Core-mmrag: Cross-source knowledge reconciliation for multimodal rag. arXiv preprint arXiv:2506.02544 (2025)
27. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
28. Wang, P., Wu, Q., Shen, C., Dick, A., Van Den Hengel, A.: Fvqa: Fact-based visual question answering. IEEE transactions on pattern analysis and machine intelligence **40**(10), 2413–2427 (2017)
29. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
30. Wu, W., Song, Z., Zhou, K., Shao, Y., Hu, Z., Huang, B.: Towards general continuous memory for vision-language models. arXiv preprint arXiv:2505.17670 (2025)
31. Yan, Y., Xie, W.: Echosight: Advancing visual-language models with wiki knowledge. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 1538–1551 (2024)
32. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
33. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5625–5644 (2024)

34. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)