TRAJECTORY ANALYSIS AND OPTIMIZATION OF AGENTIC MULTI-MODAL RAG UNDER REFERENTIAL DEFINITENESS MODULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic multimodal Retrieval-Augmented Generation (RAG) systems rely on Vision-Language Models (VLMs) to formulate retrieval queries. However, these queries often contain under-specified references to visual entities, causing the retriever to return semantically related but functionally non-resolving evidence, thereby impairing downstream multi-hop reasoning. While existing work on multimodal disambiguation largely assumes static pipelines or manually clarified references, to the best of our knowledge we provide the first systematic study of how such referential under-specification propagates through iterative retrieval trajectories and impacts final answers. To enable rigorous diagnosis, we construct a multi-dataset benchmark and introduce Retrieval-Involved Trajectory Efficiency (RTE), a metric that quantifies the novelty and utility of retrieved evidence. Controlled interventions using oracle entity grounding and entity obfuscation across various VLM backbones reveal that obfuscation increases factual novelty at the expense of utility, while oracle grounding yields diminishing returns as backbone capability grows. Notably, final answers exhibit surprising robustness to ambiguous entity references compared to their intermediate reasoning trajectories. We further identify trajectory stagnation as a recurring failure mode where naive query rewriting offers limited remedy. Motivated by these findings, we propose IIGCR, a training-free optimization mechanism that detects stagnation online and recovers degraded trajectories through context reconstruction and referent-aware rewriting. Extensive experiments demonstrate that IIGCR effectively leverages trajectory-level failure signals for online repair, consistently enhancing both final answer quality and trajectory efficiency. Our code will be publicly available upon acceptance.

1 INTRODUCTION

Recent advances in Retrieval-Augmented Generation (RAG) have led to the widespread adoption of agentic pipelines, such as MM-ReAct (Yang et al., 2023), where Vision-Language Models (VLMs) dynamically formulate retrieval queries and iteratively leverage external evidence for multi-hop reasoning (Yao et al., 2022; Li et al., 2024). Despite their flexibility, these systems expose a critical mismatch: retrieval typically operates purely over textual queries, whereas the VLM's internal reasoning is grounded in multimodal context (Liu et al., 2025; Jin et al., 2025a; Wu et al., 2026). As a result, the retriever observes only the emitted query string, with no access to the visual context needed to resolve the VLM's intended referents.

Grounded in the structural linguistics view that "the signifier's signified is stabilized by context rather than inherently guaranteed by the expression itself" (De Saussure et al., 1916), we formalize this mismatch as **referential under-specification**: the state where a textual query fails to encapsulate the multimodal context required to uniquely resolve its target entity. For instance, a VLM may emit queries such as "the sculpture in the image" or "the person associated with it" without explicitly identifying the referenced entity in a search-friendly canonical entity form. As illustrated in Figure 1(a), such under-specified queries induce **retrieval stagnation**, where semantically related yet non-resolving evidence causes repetitive retrieval trajectories. Figure 1(b) further shows that this phenomenon persists across state-of-the-art VLM backbones.

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.

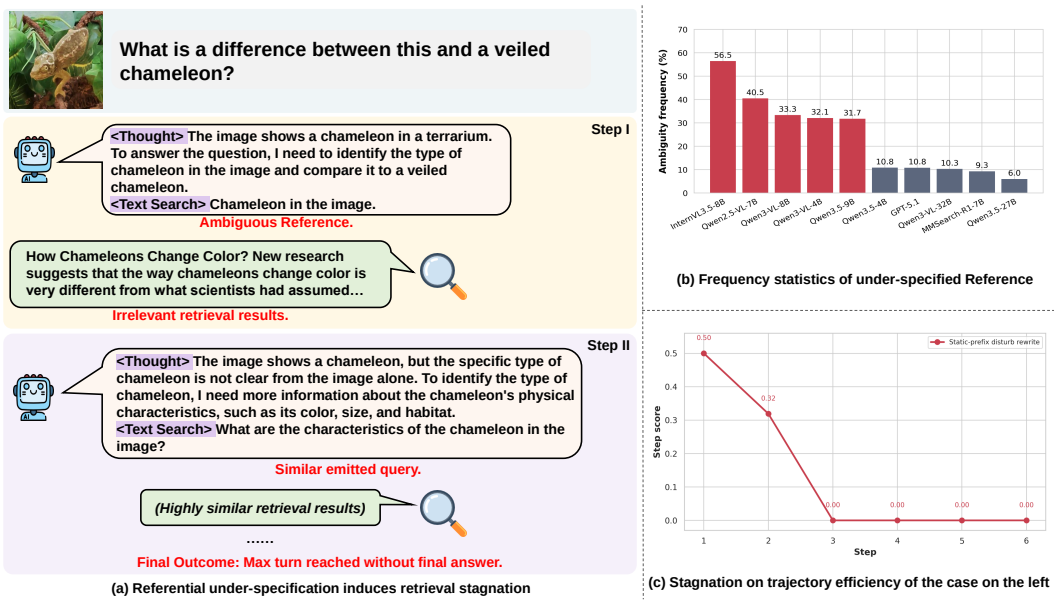

Figure 1: (a) Referential under-specification induces retrieval stagnation, where ambiguous queries yield semantically related but non-resolving evidence, leading to repetitive loops and premature termination under the interaction budget. (b) Frequency of reference under-specification across diverse backbones. (c) Step-wise trajectory efficiency S_t (defined in Section 2) demonstrating late-stage information redundancy.

In multimodal information-seeking tasks, early retrieval queries exert a compounding effect on downstream execution. Under-specified references can cause the retriever to return evidence that is semantically related to the query but insufficient to resolve the intended entity. Subsequent queries are then conditioned on this unhelpful evidence, allowing the error to propagate across later retrieval steps as shown in Figure 1(c). While existing multimodal disambiguation studies focus on static pipelines and immediate prediction quality (Jang et al., 2025; Liu et al., 2023), they treat ambiguity as a local resolution task. In contrast, we argue that referential under-specification in agentic RAG is an interface propagation failure, whose long-horizon dynamics remain under-explored.

To systematically diagnose this behavior, we treat referential definiteness as a dynamically controllable variable at the query-retriever interface. We construct the **RefAmb** benchmark for controlled diagnosis and introduce **R**etrieval-Involved **T**rajectory **E**fficiency (**RTE**), a trajectory-level metric that quantifies the novelty and utility of acquired evidence across iterative steps. Our systematic evaluation reveals three key insights: (1) Weakening referential definiteness increases factual novelty at the expense of retrieval utility; (2) Explicitly restoring under-specified entity references yields diminishing returns as model capabilities grow; and (3) Failed trajectories predominantly manifest as trajectory stagnation, a state characterized by repetitive retrieval and low information gain, where naive query rewriting offers negligible recovery.

Motivated by these findings, we propose **I**nformation **I**ncrement **G**uided **C**ontext **R**econstruction (**IIGCR**), a training-free online repair mechanism that detects stagnation during execution and restores degraded trajectories through context reconstruction and referent-aware query rewriting. Extensive experiments demonstrate that IIGCR improves both trajectory efficiency and final answer accuracy over query-rewriting baselines, with limited token overhead.

In summary, our contributions are fourfold: (1) To the best of our knowledge, we provide the first systematic study of referential under-specification as a trajectory-level interface propagation failure in agentic multimodal RAG. (2) We introduce the RefAmb diagnostic suite and RTE, a process-level metric for quantifying information novelty and utility in multi-hop retrieval. (3) We show that restoring referential definiteness yields diminishing gains as backbone capability increases, and further identify trajectory stagnation as a prevalent failure mode that naive query rewriting cannot resolve. (4) We propose IIGCR, a training-free online repair mechanism that leverages trajectory-

level failure signals to resolve stagnation, thereby improving both trajectory efficiency and final answer quality.

2 REFAMB DATASET CURATION AND RTE FOR MEASURING TRAJECTORY ACTIVITY

Task Taxonomy of RefAmb. RefAmb is a diagnostic benchmark dataset designed to study reference uncertainty in agentic RAG. It encompasses five entity-centric task categories: Entity Recognition (abbreviated as E.R.) task focuses on locating target entities. Single-hop Attribute Query (abbreviated as S.A.Q) requires extracting single-level attributes associated with target entities. Multi-hop (abbreviated as M.H.) mandates multi-step deduction relying on intermediate entities or relational links. Comparison (abbreviated as Comp.)involves conducting comparative analysis across multiple candidate objects. Complex Reasoning Pattern (abbreviated as C.R.P.) covers composite reasoning paradigms with multiple constraints and higher structural sophistication.

Data Sources of RefAmb. We sample candidate instances from OVEN (Hu et al., 2023), InfoSeek (Chen et al., 2023), CRAG-MM (Yang et al., 2024), and MC-Search (Ning et al., 2026), covering visual entity recognition, knowledge-intensive VQA, and complex multimodal reasoning. We construct RefAmb through a four-stage funnel–shaped screening pipeline. Besides, to carry out specification modulation experiments, we constructed standard entity annotations for each sample based on the semi-automatic annotation process involving LLMs. Details for data construction and reliability test are presented in Appendix C and D. The final dataset contains a total of 1500 samples, with 300 samples for each task type.

RTE Metrics. We define Retrieval-involved Trajectory Efficiency (RTE) to measure whether retrieval steps introduce novel information that is effectively used in subsequent reasoning. For each retrieval step t, the agent emits a query q_t and receives evidence r_t, from which we extract semantically deduplicated factual units F_t. Factual novelty is defined as

$$\Delta F_t = \mathbf{1}\left[\mathrm{sim}(q_t, q_{t-1}) < \tau_q\right] \cdot \frac{|F_{\leq t} \setminus F_{<t}|}{|F_t| + \epsilon}, \tag{1}$$

where $F_{<t}$ denotes previously accumulated facts. Retrieval utility $U_t \in \{0, 0.5, 1\}$ reflects whether the retrieved evidence is ignored, partially used, or fully used in the subsequent reasoning step; sensitivity to the partial-utility value is reported in Figure 7. We define the effectiveness of retrieval step (t) as

$$S_t = U_t \cdot \Delta F_t, \tag{2}$$

and compute RTE as the arithmetic mean of S_t over all retrieval steps in the trajectory. A higher RTE therefore indicates that the retrieval-feedback loop consistently acquires novel evidence while converting it into useful reasoning progress.

3 MECHANISM ANALYSIS OF REFERENTIAL UNDER-SPECIFICATION

3.1 EXPERIMENT SETUP

We implement a MM-ReAct pipeline using InternVL3.5-8B, Qwen2.5-VL-Instruct and Qwen3-VL-Instruct[1] series as the main backbones based on the FlashRAG toolkit (Jin et al., 2025b). The core implementation follows OmniSearch XML outputs (Li et al., 2024), but we re-engineered it to improve robustness. More experiment setup details are exhibited in Appendix F. To implement referential definiteness modulation, we compare the original trajectories with two minimal in-situ intervention schemes.

Oracle Intervention is designed to boost definiteness: for each original trajectory, we locate the first text retrieval query containing a core entity with ambiguous reference. We only locally rewrite this query by replacing ambiguous references with annotated oracle entities. The rewritten query is resubmitted to the retriever, the original retrieval feedback is cleared, and the agent proceeds with

[1]"-Instruct" will be omitted in the following text in Qwen2.5-VL or Qwen3-VL, but still refers to the Instruction-tuned model.

the subsequent trajectory workflow based on the newly obtained evidence. Trajectories without ambiguous references remain unchanged.

Disturbance Intervention reduces definiteness in the opposite direction: for trajectories whose initial queries contain explicit core entities, we replace such entities with their hypernyms or anaphoric expressions, update the retrieval results simultaneously, and continue running the trajectories following the same reasoning workflow. The above two methods only intervene in the first under-specified round to study the spontaneous change and potential repair capability of the Agent system.

Implementation of RTE. We use Qwen2.5-7B to extract information from retrieved documents and employ GPT-4o for Step Utility annotation. Instead of letting GPT conduct an in-depth analysis of the characterization of its internal reasoning effectiveness, we measure the explicit semantic uptake of retrieved evidence in the subsequent reasoning step to mitigate bias stemming from the model's expressive habits, illustrated in Appendix L.4. Moreover, Table 7 shows high consistency between Utility judgments, overall RTE and human evaluations, which also indirectly demonstrates the reliability of fact extraction. To ensure high diagnostic accuracy, we set $\tau_q = 1$ in Equation 1 to only exclude identical queries.

Before trajectory-level intervention analysis, we also examine when referential under-specification is most likely to arise in RefAmb. We find that it concentrates most strongly in samples where the target entity cannot be identified from text alone while external evidence is still required. This suggests that high multimodal coupling together with external-knowledge dependency is the main trigger condition for exposing under-specified references at the query–retriever interface; detailed distributions are provided in Appendix G.2.

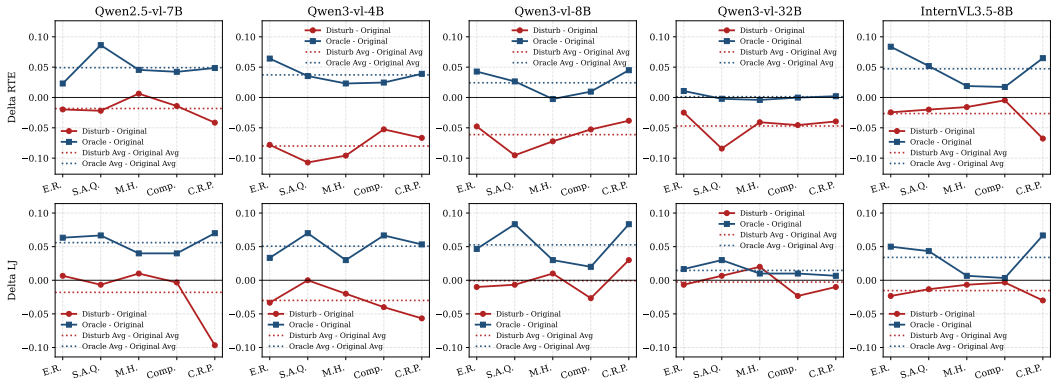

Figure 2: Cross-backbone effects of Disturbance and Oracle interventions on trajectory efficiency (RTE) and LLM-judged final answer quality (LJ). The figure records the difference of the two metrics before and after intervention. The red curve represents the changes under disturbance intervention, and the blue curve stands for Oracle. Interventions are applied only to eligible subsets, while ineligible samples retain their Original scores.

3.2 Analysis of Trajectory Robustness under Referential Definiteness

Figure 2 illustrates the impacts of two intervention methods on trajectories and final generation quality. We observe that these two types of interventions exhibit asymmetric effects: Disturbance interventions that weaken entity specification generally degrade trajectory efficiency. However, this impact does not monotonically decline with model capability across different models: even stronger models can suffer noticeable trajectory-level disruption. Additionally, generation quality also deteriorates, though stronger models demonstrate a certain degree of robustness against this degradation. This pattern is consistent with the possibility that stronger VLMs compensate for degraded retrieval through internal knowledge, visual recognition, or more effective evidence filtering, yet this does not guarantee equal stability in their retrieval trajectories.

In contrast, explicitly enhancing the specification of references generally improves trajectory efficiency and generation quality, which indicates that explicitly supplementing entity-level anchor points helps optimize retrieval alignment and boosts the utilization of evidence in subsequent rea-

soning. However, this enhancement effect **plateaus as model capability increases**. Particularly with the more powerful Qwen3-VL-32B model, the performance gains on RTE and LJ are relatively marginal. This suggests that when a model already possesses strong visual grounding, parametric knowledge, or evidence filtering capabilities, merely resolving under-specified references in queries cannot continuously yield comparable improvements in trajectory performance.

In addition, we also find that final answer quality can mask trajectory inefficiency, which is a strong motivation to come up with RTE. Table 9 reveals a moderate positive correlation between RTE and the final LLM-Judge Score. This indicates that the more new information the retrieval steps in the trajectory can provide for subsequent reasoning, the higher the quality of the final answers tends to be. Nevertheless, high-quality final answers do not entirely rely on high RTE values. To interpret the moderate RTE–LJ correlation, we evaluate Direct VLM on Low-RTE-Correct cases, where original trajectories are inefficient but still answer correctly. As shown in Table 10, Direct VLM preserves most of these cases, suggesting that **their final answers rely more on intrinsic visual understanding or parametric knowledge than on effective retrieval-feedback reasoning trajectories.** This supports our use of RTE and LJ as complementary signals for process efficiency and final-answer quality, respectively.

3.3 Internal Mechanism Analysis on Effect of Referential Definiteness

We now further decompose the internal mechanism behind the phenomenon mentioned above. Since RTE is jointly determined by step-level factual novelty and retrieval utility, pipeline inactivation is not limited to the case where the model stops acquiring new information. It can also emerge when the model continues to retrieve, but the retrieved content fails to translate into meaningful reasoning progress.

Table 1: Backbone-level metric changes relative to the Original setting. For any metric x, $\Delta x_D = x_{\text{Disturb}} - x_{\text{Original}}$ and $\Delta x_O = x_{\text{Oracle}} - x_{\text{Original}}$. The ΔF and U of each trajectory are the average values of all steps. Positive values indicate increases relative to Original, while negative values indicate decreases. This table illustrates that Disturbance boosts the novelty of retrieval yet impairs inference usability. Additionally, as the Backbone becomes stronger, the usability of Oracle rewriting gradually reaches saturation.

Model	Δ**RTE**$_D$	Δ**RTE**$_O$	$\Delta(\Delta F)_D$	$\Delta(\Delta F)_O$	ΔU_D	ΔU_O
InternVL3.5-8B	-0.027	+0.047	+0.333	+0.052	-0.112	+0.100
Qwen2.5-VL-7B	-0.018	+0.049	+0.202	+0.017	-0.065	+0.045
Qwen3-VL-4B	-0.080	+0.037	+0.336	-0.072	-0.055	+0.044
Qwen3-VL-8B	-0.061	+0.024	+0.330	-0.055	-0.045	+0.028
Qwen3-VL-32B	-0.047	+0.001	+0.418	+0.015	-0.045	-0.002

Table 1 shows that disturbance intervention increases factual novelty overall across backbones, but reduces both RTE and retrieval utility. **This reveals that negative effect of under-specification is not caused by retrieving less relevant information, but by making the retrieved information less usable.** Disturbance intervention therefore shifts trajectory inactivation from "low-novelty stagnation" to "low-utility exploration": agents continue to retrieve new facts, yet fail to convert them into effective reasoning progress. One possible reason is that replace entity names with hypernyms or anaphoric expressions broadens the semantics of queries, thereby improving the overall retrieval breadth of the trajectory.

Oracle behaves more selectively. For models with relatively weak capabilities, Oracle improves both RTE and utility, showing that canonical entities provide stable retrieval anchors and help suppress low-yield retrieval states. However, this effect diminishes as backbone capability increases. On the Qwen3-VL series, oracle remains close to the Original setting across RTE, ΔF, and even causes a slight drop in Utility, indicating a saturation regime in which the model already has sufficient entity localization or evidence-selection ability, leaving limited room for oracle-level clarification to further improve trajectory dynamics.

3.4 TRAJECTORY STAGNATION: INTERNAL SYMPTOMS OF THE FAILURE TRAJECTORY

We further analyze failed trajectories from a temporal perspective by focusing on Original MTR trajectories, i.e. those trajectories that reach the maximum retrieval budget but still do not end in `OK`. As exhibited in Figure 3, these long failed trajectories exhibit a **consistent low-gain pattern across backbones**: both factual novelty and the effective step score S_t decrease in later iterations, while the cumulative effective contribution gradually levels off. This shows that the retrieval-feedback loop continues to run but contributes little additional useful evidence for reasoning. We refer to this phenomenon as **trajectory stagnation**. Based on this phenomenon, we further analyze the impact of referential definiteness on trajectory state transitions and trajectory robustness. Specifically, we construct trajectory state transition matrices for subsets of samples that undergo state changes under the two intervention conditions of Oracle and Disturbance interventions respectively.

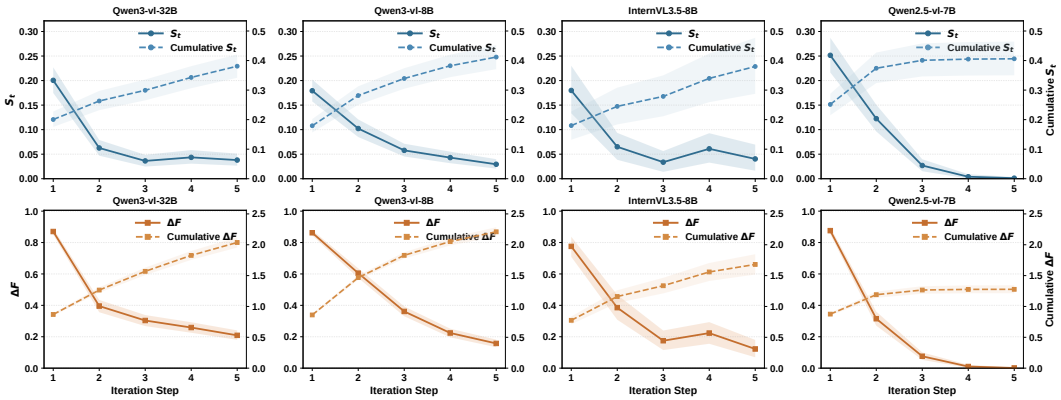

Figure 3: Trajectory stagnation in Original long-horizon failures. For each backbone, we report the step-wise effective progress S_t on the left y-axis and its cumulative gain on the right y-axis, using trajectories that do not reach `OK` while exhausting the maximum retrieval budget.

As shown in Figure 4, Oracle and Disturbance interventions lead to different transition patterns for original MTR trajectories. Oracle yields a higher MTR-to-OK transition rate and a lower MTR-to-MTR persistence rate, indicating that improved referential definiteness can partially restore stagnated trajectories. In contrast, Disturbance increases MTR persistence and reduces MTR-to-OK recovery, suggesting that degraded referential definiteness makes stagnated trajectories harder to restore in subsequent iterations.

Nevertheless, Oracle does not fully restore trajectory robustness. Some original MTR trajectories still remain non-OK after intervention, and a non-negligible fraction of original OK trajectories regress to MTR or GE. These results indicate that explicit referential information improves the input condition but does not guarantee stable reasoning or active self-correction. Thus, oracle intervention provides only limited mitigation, while disturbance mainly weakens robustness by reducing the recoverability of MTR trajectories.

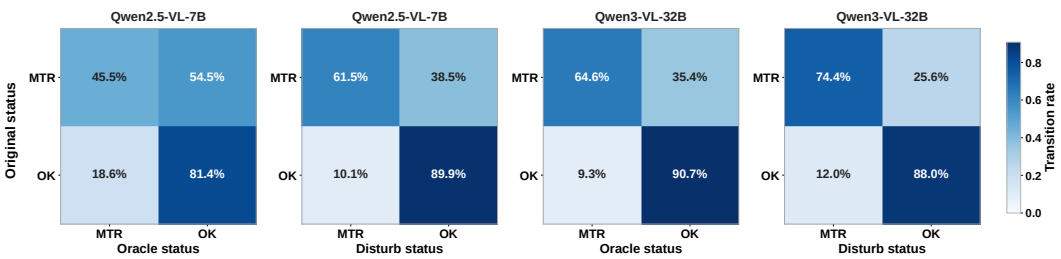

Figure 4: Trajectory state transition matrices under the modulation of different referential definiteness. All results are statistically calculated on subsets of samples subjected to oracle or disturbance interventions. MTR means reaching the iteration quota without obtaining the final answer, while OK indicates successfully yielding the final answer.

4 FROM DIAGNOSIS TO REPAIR: INFORMATION-INCREMENT-GUIDED CONTEXT RECONSTRUCTION

The above analysis shows that referential under-specification affects both retrieval trajectories and final generation quality by altering information increment and evidence utility. Although oracle entity anchoring can partially mitigate these failures, it provides only an idealized one-step correction, as entity grounding in real scenarios typically requires iterative evidence acquisition and retrieval replanning. We therefore shift our entry point to trajectory stagnation, the observable execution-level consequence of referential under-specification. This motivates **I**nformation **I**ncrement **G**uided **C**ontext **R**econstruction (IIGCR), which uses stagnation signals to prompt the VLM to reassess its current state and replan subsequent retrieval actions. Through IIGCR, we test whether the diagnosed trajectory symptoms can serve as actionable online repair signals while reducing unproductive token consumption without degrading final-answer quality. The overall framework is shown in Figure 5 and detailed algorithms are provided in Appendix H.

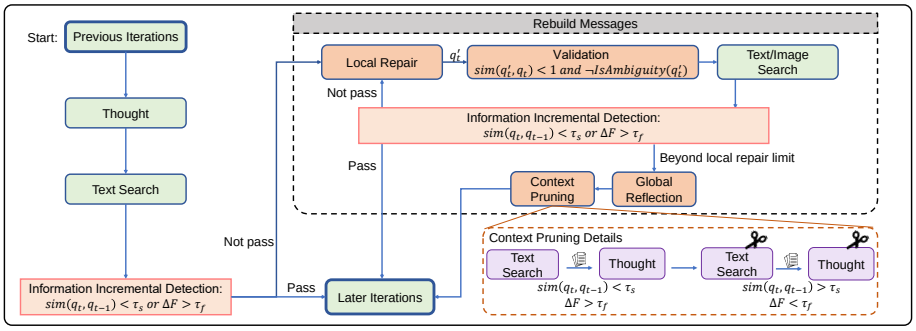

Figure 5: Overview of IIGCR: an online repair framework for detecting poor retrieval states and repairing them through context reconstruction and adaptive query revision.

Hierarchical Stagnation Detection and Repair. As shown in Figure 5, IIGCR follows a coarse-to-fine repair hierarchy that aims to maximally preserve the integrity of the spontaneous reasoning trajectories of VLMs. It first uses two lightweight gates to identify inactive retrieval-feedback loops. Before retrieval, a repetition gate is triggered when $\mathrm{sim}(q_t, q_{t-1}) > \tau_s$; if the query passes this check, a post-retrieval low-yield gate is applied when $\Delta F_t < \tau_f$. The online estimate of ΔF_t is obtained by chunking retrieved results and semantically deduplicating them by cosine similarity, avoiding the heavier factual-unit extraction used in offline RTE evaluation.

Once either gate is activated, IIGCR first performs a **local repair**, since many failures can be corrected by revising only the current retrieval action. The VLM is prompted to identify the missing clue—such as an entity, attribute, relation, temporal cue, or spatial cue—and accordingly revise the text query, switch to `Image Retrieval`, or terminate retrieval with `No Retrieval`. The rebuilt action is then **validated** to reject near-duplicate rewrites. This local stage therefore addresses immediate query-level defects while preserving as much of the existing trajectory as possible.

Visual Anchoring for Unresolved Referents. When local repair identifies that the missing clue is visually grounded and cannot be recovered through textual reformulation alone, IIGCR escalates to visual anchoring. The VLM generates a short description of the required visual evidence, which is used by Grounding DINO (Liu et al., 2024) to localize and crop the relevant region of interest (ROI) from the original image. The resulting ROI is then supplied to image retrieval. In this way, visual anchoring complements local rewriting by providing an explicit visual entity anchor when lexical revision alone is insufficient, without relying on oracle entity names.

Trajectory-Level Recovery through Reflection and Context Pruning. If low-gain behavior still occurs after local repair, manifesting that local repair has been performed the specified maximum number of times, the failure is treated as no longer local but trajectory-level: ineffective retrieval states have accumulated and may continue to bias subsequent reasoning. IIGCR therefore switches to **global reflection**, prompting the VLM to reassess the current reasoning state and choose a genuinely different retrieval plan. At the same time, the dialogue history is reconstructed through prun-

ing. Rather than replacing only a single <Search></Search> span, we model the trajectory as a stack and remove low-contribution or duplicated retrieval states together with their adjacent feedback segments from back to front. In this way, reflection changes the future plan, while pruning removes the historical evidence that sustains the failed plan. Their combination allows inference to resume from a cleaner context with a revised retrieval strategy. Detailed prompts and algorithms are provided in Appendix L.8.

Table 2: Main results of IIGCR and baselines across task types in RefAmb. $\tau_f = 0.20$ and $\tau_s = 0.90$. **Bold** denotes the best value within each backbone group, and underlining denotes the second-best value. "$*$" and "$***$" denotes two-sided t-test with $p < 0.05$ and $p < 0.001$ over the best baseline.

Backbone	Method	E.R.		S.A.Q.		M.H.		Comp.		C.R.P.		Avg.	
		RTE ↑	LJ ↑	RTE ↑	LJ ↑	RTE ↑	LJ ↑	RTE ↑	LJ ↑	RTE ↑	LJ ↑	RTE ↑	LJ ↑
Qwen2.5-VL-7B	ReAct	20.78	38.00	27.13	13.00	29.96	12.00	28.26	15.67	49.57	49.67	31.14	25.67
	MMSearch-R1	12.90	**48.33**	29.66	**26.33**	25.69	11.67	27.27	20.00	24.80	52.00	24.06	31.67
	Prompt Constraint	19.02	32.00	26.89	12.00	34.66	13.33	26.75	16.33	47.25	50.67	30.91	24.87
	Ambiguity Rewrite	20.31	32.67	30.61	14.67	33.13	13.00	26.74	15.67	48.00	50.67	31.76	25.33
	Oracle Rewrite	23.08	44.33	35.75	19.67	34.52	16.00	32.49	19.67	**54.43**	**56.67**	36.05	31.27
	IIGCR	**28.67**	47.00	**40.03**	20.67	**42.40**	**19.33**	**38.25**	**26.67**	45.09	53.67	**38.89*****	**33.47***
Qwen3-VL-32B	ReAct	31.32	48.67	**43.54**	31.00	30.44	21.33	25.31	29.33	35.01	58.33	33.12	37.73
	Prompt Constraint	24.56	42.33	37.70	30.00	27.57	21.74	19.99	24.33	**46.82**	70.92	31.33	37.86
	Ambiguity Rewrite	25.02	45.33	39.32	32.67	29.16	21.00	24.92	30.33	46.36	72.15	32.95	40.30
	Oracle Rewrite	**32.38**	50.33	43.32	34.00	30.04	22.33	25.29	30.33	35.21	59.00	33.25	39.20
	IIGCR	31.39	**53.33**	43.36	**36.00**	**34.54**	**29.33**	**26.22**	**35.33**	42.81	**73.00**	**35.66*****	**45.40*****
GPT-5.1	ReAct	22.46	49.33	28.76	41.67	21.27	32.33	19.94	46.00	28.15	84.33	24.12	50.73
	Oracle Rewrite	22.44	**49.67**	29.00	**43.00**	21.69	33.00	20.08	**46.67**	28.44	**85.67**	24.33	**51.60**
	IIGCR	**22.85**	49.33	**29.17**	41.33	**22.55**	**33.67**	**20.61**	45.67	**34.09**	**85.67**	**25.85***	51.13

4.1 Main Results

Baselines. Prompt Constraint guides VLMs to proactively avoid generating ambiguous queries by adding a rule to the ReAct prompt: "Do not generate queries with vague references; if the entity name cannot be confirmed, perform image retrieval first." **Ambiguity Rewrite** activates under-specified reference discrimination when invoking Text Retrieval. For ambiguous references, it retrieves entity names via a single image retrieval call with ROI-clipped image and rewrites the queries to eliminate referential ambiguity.

Table 3: IIGCR and oracle intervention effects on the intervention subset. ΔOK denotes the change in the proportion of samples with OK trajectories relative to Original, and ΔNotOK denotes the change in the proportions of samples with status ≠ OK and reach the max turns limit.

Intervention	ΔOK	ΔNotOK
Qwen2.5-VL-7B		
Oracle	+20.6	-18.4
IIGCR	**+75.2**	**-43.2**
Qwen3-VL-32B		
Oracle	+11.8	-41.2
IIGCR	**+68.0**	**-68.0**

Table 2 summarizes the overall performance on RefAmb. Compared with standard ReAct, IIGCR consistently improves RTE across all three backbones, demonstrating the effectiveness of online trajectory repair in reducing retrieval inefficiency. It also improves LJ on Qwen-VL models while maintaining comparable answer quality on GPT-5.1. On Qwen2.5-VL-7B, IIGCR achieves gains of 7.75 RTE points (+24.9%) and 7.80 LJ points (+30.4%) over ReAct, with positive but smaller improvements on stronger backbones. This suggests that trajectory-level intervention remains beneficial even when VLMs become stronger, although stronger models can partially compensate for inefficient retrieval through their intrinsic abilities. Notably, without additional training, IIGCR outperforms the RL-tuned MMSearch-R1 baseline by 14.83 RTE points and 1.8 LJ points. This is because MMSearch-R1 is trained on when to search, how to search, and when to stop, but does not take the feedback from retrieval results and gain detection for the entire trajectory into account.

The limited gains from oracle entity anchoring further reveal the advantage of trajectory-level intervention. As backbone capability increases, the benefit of explicitly providing entity anchors gradually diminishes, suggesting that stronger VLMs already possess improved visual localization and evidence filtering abilities. In contrast, IIGCR remains effective because it does not rely on explicit

entity availability, but instead exploits universal trajectory-level failure signals to adaptively repair retrieval behaviors.

Prompt Constraint slightly degrades performance on smaller models, while providing marginal improvements in final-answer correctness on stronger models. This observation suggests that referential definiteness reflects a fundamental information asymmetry between VLMs and retrievers. Thus, smaller models may lack the capability to explicitly recognize and compensate for this mismatch through prompt-level constraints. Furthermore, repeated Ambiguity Rewrite only achieves noticeable gains on Qwen3-VL unexpectedly, indicating that the effectiveness of online ambiguity intervention is heuristic, unstable and highly model-dependent.

Further Analysis. Table 4 compares the inference overhead of Original and IIGCR. IIGCR slightly increases average latency due to additional repair and rewriting, but reduces the average total token consumption on both backbones. The reduction mainly comes from fewer input tokens, as IIGCR prunes redundant or low-gain trajectory segments instead of carrying the full accumulated ReAct history. Thus, IIGCR achieves better token efficiency with moderate runtime overhead, especially on weaker backbones.

Table 4: Comparison of VLM calling costs and latency between original and IIGCR across Qwen2.5-VL-7B and Qwen3-VL-32B.

Model	Method	Avg. Input Tokens	Avg. Output Tokens	Avg. Total Tokens	Avg. Latency (s)
Qwen2.5-VL-7B	Original	23031.23	285.80	23317.03	7.90
Qwen2.5-VL-7B	IIGCR	20064.53	338.30	20402.83	9.54
Qwen3-VL-32B	Original	8456.81	460.65	8917.46	52.17
Qwen3-VL-32B	IIGCR	8294.96	576.96	8871.92	56.56

Table 3 and Figure 13 show that IIGCR directly mitigates the stagnation phenomenon. Ablation study shown in Table 13 further analyzes the contribution of each component, showing that the modules are complementary: local repair improves immediate low-yield search actions, visual redirection handles cases where the missing clue is visually grounded, and global reflection with context pruning becomes important when inefficient retrieval states have accumulated in the dialogue history. Figure 11 and Figure 12 report the sensitivity analysis over the two core triggering thresholds. IIGCR remains robust across a reasonable range of ΔF_t and query-similarity thresholds, suggesting that the method is not overly dependent on a narrow hyperparameter choice.

5 Related Works

Recent studies have investigated how to improve multimodal retrieval by resolving missing visual references and strengthening entity-level alignment. (Liu et al., 2025) study visually under-specified retrieval queries and rewrite them with additional visual descriptions to improve retrieval effectiveness. Beyond query rewriting, multimodal RAG approaches construct explicit entity-level structures: (Yuan et al., 2025) introduce multimodal knowledge graphs to connect visual entities with external knowledge for retrieval, while (Park et al., 2026) extend this idea to multi-hop multimodal reasoning through grounded entity and relation retrieval. For agentic settings, (Zhang et al., 2026) investigate visual query preprocessing before retrieval to reduce retrieval failures. However, these studies primarily optimize retrieval quality or query representations, while we investigate how failures in visually grounded references propagate through iterative retrieval-feedback trajectories and lead to agent-level stagnation. More related works are presented in Appendix A.

6 Conclusion

The effect of referential under-specification in agentic multimodal RAG varies with both backbones and tasks: weakened queries tend to increase factual novelty while reducing retrieval utility, whereas oracle grounding benefits models unevenly and gradually saturates. These findings inspire us: (1) referential under-specification should not be evaluated only by final answer quality, because strong VLMs may mask inefficient retrieval trajectories. (2) Effective trajectory repair should be triggered by process symptoms such as repetition and low information gain, not by one-shot query clarification alone.

7 AI USE STATEMENT

In this paper, a semi-automatic annotation method involving LLMs is adopted during the data annotation process. All relevant procedures are described in Appendix C, and the consistency evaluation against human annotators is reported in detail in Appendix D to verify evaluation reliability. In addition, only GPT-5.6 is used for paper polishing and assisting in experimental code review. No AI tools are involved in experimental design, data analysis or conclusion drawing. The polished initial draft has undergone rigorous human review and extensive revisions.

8 REPRODUCIBILITY STATEMENT

The dataset construction process, theoretical derivations, etc. in this paper are introduced in detail in the appendix, serving as an important part of the integrity of our paper. The core code implementation of our research project, including the hyperparameter settings for experimental environments, implementations of RTE, baselines, IIGCR and referential definiteness modulation, as well as the required prompts, are provided in the supplementary materials. The open-source code repository will be made available if the paper is accepted.

REFERENCES

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16776–16809, 2025.

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

Longfeng Chen, Zheng Xiao, Juyuan Wang, Zeyu Huang, Yawen Zeng, and Jin Xu. HEAR: A holistic extraction and agentic reasoning framework for document understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 14376–14382, 2025.

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

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Ferdinand De Saussure et al. Nature of the linguistic sign. *Course in general linguistics*, 1:65–70, 1916.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12065–12075, 2023.

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Youngjoon Jang, Seongtae Hong, Junyoung Son, Sungjin Park, Chanjun Park, and Heui-Seok Lim. From ambiguity to accuracy: The transformative effect of coreference resolution on retrieval-augmented generation systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 422–433, 2025.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025a.

Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. FlashRAG: A modular toolkit for efficient retrieval-augmented generation research. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 737–740, 2025b.

Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Philip S Yu, Fei Huang, et al. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*, 2024.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 790–807, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding Dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Szu-Ting Liu, Wen-Yu Cho, Hsin-Wei Wang, and Berlin Chen. Bridging underspecified queries and multimodal retrieval: A two-stage query rewriting approach. In Kai-Wei Chang, Ke-Han Lu, Chih-Kai Yang, Zhi-Rui Tam, Wen-Yu Chang, and Chung-Che Wang (eds.), *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pp. 63–70, National Taiwan University, Taipei City, Taiwan, 2025. Association for Computational Linguistics.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*, 2025.

Xuying Ning, Dongqi Fu, Tianxin Wei, Mengting Ai, Jiaru Zou, Ting-Wei Li, Hanghang Tong, Yada Zhu, Hendrik Hamann, and Jingrui He. MC-Search: Evaluating and enhancing multimodal agentic search with structured long reasoning chains. *arXiv preprint arXiv:2603.00873*, 2026.

Hyeongcheol Park, Jiyoung Seo, Jaewon Mun, Hogun Park, Wonmin Byeon, Sung June Kim, Hyeonsoo Im, JeungSub Lee, and Sangpil Kim. M^3KG-RAG: Multi-hop multimodal knowledge graph-enhanced retrieval-augmented generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14968–14978, 2026.

Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Jin Di, Yu Cheng, Qifan Wang, and Lifu Huang. Rora-VLM: Robust retrieval-augmented vision language models. *arXiv preprint arXiv:2410.08876*, 2024.

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

Yubo Sun, Chunyi Peng, Yukun Yan, Shi Yu, Zhenghao Liu, Chi Chen, Zhiyuan Liu, and Maosong Sun. VisRAG 2.0: Evidence-guided multi-image reasoning in visual retrieval-augmented generation. *arXiv preprint arXiv:2510.09733*, 2025.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. VidoRAG: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9124–9145, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun MA, and Ziwei Liu. MMSearch-r1: Incentivizing LMMs to search. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2456–2487. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.acl-long.114. URL https://aclanthology.org/2026.acl-long.114/.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. CRAG - comprehensive rag benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 10470–10490. Curran Associates, Inc., 2024. doi: 10.52202/079017-0335. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1435d2d0fca85a84d83ddcb754f58c29-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-ReAct: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, volume 2025, pp. 21074–21098, 2025a.

Xiaohan Yu, Zhihan Yang, and Chong Chen. CogPlanner: Unveiling the potential of agentic multimodal retrieval augmented generation with planning. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 2–11, 2025b.

Xu Yuan, Liangbo Ning, Wenqi Fan, and Qing Li. mKG-RAG: Multimodal knowledge graph-enhanced rag for visual question answering. *arXiv preprint arXiv:2508.05318*, 2025.

Jiankun Zhang, Shenglai Zeng, Kai Guo, Xinnan Dai, Hui Liu, Jiliang Tang, and Yi Chang. Fix Before Search: Benchmarking agentic query visual pre-processing in multimodal retrieval-augmented generation. *arXiv preprint arXiv:2602.13179*, 2026.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, et al. MR^{2}AG: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *arXiv preprint arXiv:2411.15041*, 2024.

Appendix

CONTENTS

A Related Works **15**
A.1 Visual Question Answering 15
A.2 Agentic Retrieval Augmented Generation 15

B Problem Formulation **15**
B.1 Agentic Multimodal Retrieval 15
B.2 Referential Under-specification 15

C Details on Dataset Sources **16**
C.1 From Four Raw Datasets to RefAmb via Slot Filling 16
C.2 Referential Dependency Annotation and Analysis on RefAmb 17

D LLM Annotation Consistency with Human Annotations **17**

E Details on RTE **19**
E.1 Partial-Contribution Label Value Sensitivity 19
E.2 Cross Model Sensitivity 19
E.3 Correlation between RTE and Generation Quality 19

F Experiment Setup Details **20**

G More Findings **21**
G.1 Entity Ambiguity Frequency across Different VLM Backbones 21
G.2 Analysis of Trigger Conditions for Under-specified Reference 21
G.3 Trajectory Length Distribution 21
G.4 Residual Ambiguity after Oracle Intervention 22
G.5 Details on RTE and LJ across Baselines under Two Interventions 23

H Algorithms of IIGCR **23**

I Additional Analysis of IIGCR **25**
I.1 Hyper-Parameter Sensitivity 25
I.2 Ablation Study 26

J Theoretical Analysis on RTE **27**

K Case Study **27**

L Prompt Templates in the Whole Research Project **29**

L.1 Prompt for Slot Filling in RefAmb Curating 29

L.2 Prompts for Referential Dependency Annotation 30

L.3 Prompt for Ambiguous Reference Recognition 32

L.4 Prompt for Iteration Utility Annotation 32

L.5 Prompt for LLM as Judge 32

L.6 Prompt for VLM Executing ReAct 33

L.7 Prompts for Two Interventions 33

L.8 Prompt for IIGCR 34

M Limitations & Future Work **35**

A Related Works

A.1 Visual Question Answering

Visual Question Answering (VQA) was formally introduced by (Antol et al., 2015), to evaluate whether models can answer natural-language questions grounded in visual inputs. Given an image and a question referring to its visual content, a VQA system is expected to infer the relevant visual semantics and return a correct answer. Over time, VQA has expanded beyond early single-image, classification-style benchmarks to encompass compositional reasoning, external knowledge acquisition, and instruction-following scenarios (Antol et al., 2015; Hudson & Manning, 2019). Classical VQA datasets mainly emphasize perceptual recognition and short-answer supervision, while newer datasets stress long-tail entities, multi-hop reasoning, and external-knowledge dependence (Marino et al., 2019; Schwenk et al., 2022).

Methodologically, the field has progressed through several stages. Early systems predominantly followed CNN–RNN style answer classification pipelines, which were effective for simple recognition but weak at fine-grained grounding and compositional reasoning (Kazemi & Elqursh, 2017). Attention and multimodal fusion methods then improved region–question alignment, yet still relied heavily on static parametric memory (Lu et al., 2016). Large-scale pretraining and transformer-based VLMs further improved generalization and reasoning breadth, but exposed persistent issues in factual reliability and knowledge freshness (Chen et al., 2020). To address these limits, VQA has increasingly moved toward retrieval-augmented paradigms that ground generation in external evidence, and more recently toward agentic workflows with iterative planning, query refinement, and tool-assisted retrieval control (Abootorabi et al., 2025; Mei et al., 2025).

A.2 Agentic Retrieval Augmented Generation

Agentic RAG extends retrieve-then-read pipelines by letting the model plan sub-queries, trigger retrieval, and invoke tools inside a trajectory-conditioned loop. In visual retrieval more broadly, systems such as (Yu et al., 2025a; Sun et al., 2025) focus on vision-native retrieval and indexing rather than agentic planning. By contrast, agentic multimodal methods such as MM-ReAct based pipeline (Yang et al., 2023), (Li et al., 2024), planning agent like CogPlanner (Yu et al., 2025b), multi-agent paradigm like ViDoRAG (Wang et al., 2025a), and HEAR (Chen et al., 2025) move toward iterative planning and retrieval control, while (Zhang et al., 2024) and (Qi et al., 2024) further explore retrieval-reflection and robustness under complex visual evidence.

Compared with these methods, our focus is narrower: we study the impact of queries whose entity references are under-specified in VLM-driven retrieval pipelines, where the model must call retrieval tools and generate queries on its own, and analyze how such queries affect trajectory efficiency and generation quality. This is a diagnostic and mechanistic study whose findings motivate an auxiliary intervention for low-gain trajectories.

B Problem Formulation

B.1 Agentic Multimodal Retrieval

Given an image I and a question x, an agentic multimodal retrieval system produces a trajectory

$$\tau = (a_1, o_1, a_2, o_2, \ldots, a_T, y), \tag{3}$$

where each action a_t may be a thought, a sub-question, a text retrieval query, an image retrieval query or no retrieval. o_t is the feedback generated by the VLM for each action. For text retrieval actions, the agent emits a query q_t, receives retrieved evidence o_t, and conditions later reasoning on the accumulated context. y is the terminal state where the final answer is obtained upon iteration completion, which means the system succeeds only when the trajectory both acquires relevant evidence and converts it into the final answer y.

B.2 Referential Under-specification

Let $s_r(q_t)$ denote the query expression emitted by the agent and let $e^\star$ denote the intended task entity. We define the candidate referent set induced by the retriever-facing query as $\mathcal{E}(q_t)$. A query

is referentially under-specified when it does not uniquely identify the intended referent:

$$|\mathcal{E}(q_t)| > 1 \quad \text{or} \quad e^\star \notin \arg\max_{e \in \mathcal{E}(q_t)} P(e \mid q_t). \tag{4}$$

A phrase can be understandable to a human who sees the image, yet under-specified to a text retriever that lacks visual context. We therefore judge under-specification at the query-retriever interface.

C Details on Dataset Sources

C.1 From Four Raw Datasets to RefAmb via Slot Filling

I. Collection of multi-source candidate samples. While selecting visual recognition tasks OVEN (Hu et al., 2023) and classic VQA datasets InfoSeek (Chen et al., 2023), we also incorporate datasets designed for complex scenarios and reasoning paradigms, including CRAG-MM (Yang et al., 2024) and MC-Search (Ning et al., 2026). The accessible source pools and their selected sets are summarized in Table 5.

Table 5: Accessible source pools to construct RefAmb datasets.

Dataset	Scale	Selection Sets
CRAG-MM	2,621	Validation and Public test split.
MC-Search	1,469	Text-Initiated Chain and Image-Initiated Chain.
InfoSeek	73,620	Validation split.
OVEN	3,291	Query Validation split.

II. Task Slot Filling. Instead of uniformly sampling from the concatenation of all raw corpora, we first identify which task slots each source can reliably support, and then select samples to cover five diagnostic slots: **Entity Recognition**, **Single-Hop Attribute Query**, **Multi-Hop Reasoning**, **Comparison**, and **Complex Reasoning Pattern**. This design avoids over-representing large but structurally homogeneous sources, while preserving smaller sources that provide diagnostic structures unavailable elsewhere. Specifically, we classify the corresponding samples into five task slots by combining the prior information of original labels and categorization from VLMs along with manual review. For data sources with clearly defined original labels or datasets, we conduct direct mapping: OVEN is categorized into the entity recognition task; Text-Initiated Chain and Image-Initiated Chain in MC-Search are classified as complex reasoning tasks; Multi-Hop, Comparison and simple-recognition in CRAG-MM are mapped to the multi-hop, comparison and entity recognition tasks of RefAmb respectively. For InfoSeek and other categories in CRAG-MM that cannot be directly mapped, we leverage GPT-4o for auxiliary classification and revise the auto-filled results via manual review. Considering that CRAG-MM has a relatively moderate scale yet rich diversity of types, when constructing the initial candidate pool, we adopt a random sampling method that takes the scale of CRAG-MM into account while keeping the size of the final dataset moderate. This source-aware slot filling yields a 2,803-sample RefAmb origin pool. The consistency statistics with human are provided in Appendix L.1 and Table 7. The complete prompts are provided in Appendix L.1.

III. Oracle entity labeling and filtering. We then apply three successive normalization steps, as shown in Table 6. First, we perform oracle-entity annotation on the 2,803-sample origin pool and filter out unstable cases with no entities or excessive entities (The number of extracted entities ≥ 3).

IV. Data source balancing and stratified sampling. We perform source-level re-balancing to prevent any single source from dominating the diagnostic pool after oracle filtering. Finally, for the main controlled analysis, we construct a task-balanced subset by sampling 300 instances from each of the five task slots, yielding 1,500 samples.

To assess the reliability of this source-aware normalization, we stratify samples by source–slot combination and manually inspect four types of annotations: task-slot mapping and oracle entity. For the

Table 6: Construction flow from the RefAmb origin pool to the task-balanced analysis subset.

Stage	Count	Explanation
RefAmb origin pool	2803	Source-aware slot-filled pool after cross-source merging.
Oracle-valid pool	2777	Remove samples without a stable core entity anchor, including cases with overly diffuse oracle-entity supervision.
Source-rebalanced pool	2323	Re-balance source contributions to avoid source dominance after oracle filtering.
Task-balanced subset	1500	Sample 300 instances from each of the five task slots.

binary under-specification dimensions, we report accuracy and agreement statistics; for oracle entities, we report exact-match and semantic-consistency rates. The results are shown in Appendix D.

C.2 Referential Dependency Annotation and Analysis on RefAmb

To characterize the problem of ambiguous reference, we label each sample with two binary diagnostic variables: C_1: Text Alone Identifies Entity and C_2: External Evidence Dependency. C_1 measures whether the target entity can be located solely through text queries. If lexical cues in the question can uniquely identify the entity without image context, the sample is marked as $C_1 = \text{Yes}$. C_2 measures whether external evidence beyond the image and question text is required to answer the query. We obtain this label via a two-stage pipeline: (1) generate detailed descriptions of images; (2) leverage a larger-scale model to determine whether the question can be answered using only these descriptions and the original question text.

We calculated the joint distribution of C_1 and C_2 across different task types in the RefAmb dataset, shown in Figure 6. We argue that a high marginal probability for $C_1 = \text{No}$ indicates that entity anchoring relies heavily on images, i.e., **high modality coupling**. In such cases, a high marginal probability for $C_2 = \text{Yes}$ means the retrieval tool is more likely to be invoked multiple times, raising the probability of under-specified references in the generated queries. In this case, the heatmap blocks corresponding to $C_1 = \text{No}, C_2 = \text{Yes}$ appear in a darker color. Entity Recognition is the tightly image-coupled family, so the query often cannot stand on its own and the image is doing most of the work in resolving both C_1 and C_2. Single-Hop Attribute Query and Multi-Hop are different in form but similar in mechanism: they ask for attributes, relations, or derived facts that are not fully encoded in the surface wording, so the text often leaves the entity under-specified while the answer still depends on outside knowledge. Comparison tasks are structurally mixed, because one side of the comparison may already be explicit while the other still needs to be inferred. That is why both C_1 outcomes naturally coexist there. Complex Reasoning Pattern is driven by radiated sub-problems around a core entity, and in MC-Search, the coupling degree between the two modalities of the problem is relatively low, so the table reflects a much stronger reliance on text-side entity anchoring than on visual disambiguation alone. Table 11 shows that such problems have a lower probability of referential under-specification, indicating that the occurrence of referential under-specification is potentially correlated with the modal coupling and retrieval dependency of Queries.

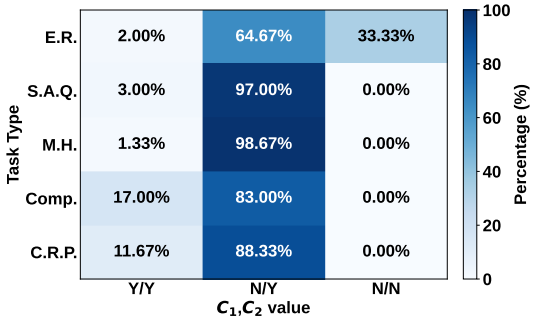

Figure 6: Per-task-type distribution on C_1 and C_2 in RefAmb, where each cell is the corresponding labels of C_1 and C_2.

D LLM Annotation Consistency with Human Annotations

We use GPT-4o to annotate the referential definiteness of queries. Therefore, we recruited a PhD student specializing in the same field to manually annotate a total of 250 sampled queries extracted

from the ReAct trajectories of various models, and the κ coefficient was subsequently calculated as shown in Table 7, which indicates a high degree of consistency between the two sets of annotations.

We also employed GPT-4o for annotation during the construction of RefAmb, followed by manual review to categorize tasks for the InfoSeek and CRAG-MM subsets. Given the extremely imbalanced label distribution of this task, the κ value is inherently low. To quantify consistency more accurately, we adopted Gwet's AC1 to rectify the bias of κ.

Table 7: Alignment between human judgments, and GPT labeling. Trajectory-level alignment is measured by rank correlation between human overall labels and RTE. Step-level alignment is measured by binarizing human usefulness labels and model utility scores on all shared annotated steps. Gwet's AC1 is adopted to estimate consistency for both label distribution skewness $C_{1,2}$

Evaluation Level	Metric	Value
Referential Under-specification Label		
Binary Label	Cohen's κ	0.8225
RefAmb Dataset Label		
Task Slot Filling	Gwet's AC1	0.9899
Oracle Entity Consistency	EM	0.8300
Oracle Entity Consistency	Entailment Acc	0.9500
C_1 Consistency	Gwet's AC1	0.9319
C_2 Consistency	Gwet's AC1	0.9233
RTE Label		
Trajectory-level Quality	Spearman's ρ	0.7061
Trajectory-level Quality	Kendall's τ	0.6467
Step-level Utility	Cohen's κ	0.7198

For Oracle Entity, we provide human annotators with the same information available to the large model: the context consists of questions, images, answers from the source dataset, as well as intermediate trace information (if any). Human annotators are then asked to infer core entities. Finally, we calculate the EM and Entailment Accuracy of GPT-4o's annotations against human annotations. The results are presented in Table 7, demonstrating a high level of consistency.

We also employed GPT-4o to annotate two labels $C_{1,2}$ for referential dependency. The C_2 label adopts a two-stage annotation pipeline: first, a small model (Qwen2.5-VL-7B) generates a purely superficial description of the image, followed by a judgment of whether the question can be directly answered based on this description. We do not ignore the risk that inaccurate image descriptions may trigger error propagation and degrade the accuracy of C_2. To address this issue, we further conducted human inter-annotator agreement evaluation. We sampled 300 data instances from RefAmb for consistency assessment. The final results are presented in Table 7, which demonstrates a high level of inter-annotator consistency.

Similarly, in the calculations of RTE, we also utilize Qwen2.5-7B for fact extraction and GPT-4o for utility inference. We recognize that small-scale models introduce errors during fact extraction. Even though we reformulate the utility task into a semantic comprehension task, biases inherent to model preferences may still induce errors. To address this issue, we perform consistency estimation. Specifically, human evaluators are asked to assign identical scores to both utility performance and the overall trajectory quality. A high degree of consistency between the two metrics allows us to reasonably conclude that the overall reliability of fact extraction is sufficient. The final results are presented in Table 7, demonstrating acceptable consistency for both metrics.

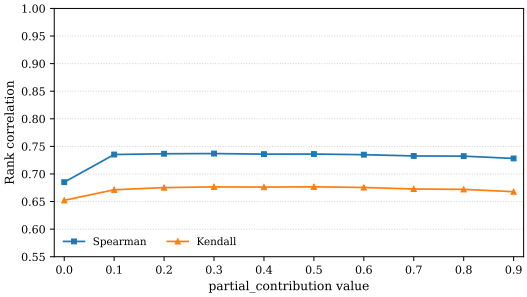

Figure 7: Hyper-parameters sensitive analysis for partial-contribution label value in the calculation of RTE.

E Details on RTE

E.1 Partial-Contribution Label Value Sensitivity

We conduct hyper-parameter sensitivity analysis for partial-contribution label values in the calculation of U_t within the RTE system, with values ranging from 0.1 to 0.9. Subsequently, we estimate the consistency between the RTE values obtained from different partial values and human evaluations of trajectory quality, using Spearman and Kendall correlation coefficients to quantify the degree of consistency. The results are shown in Figure 7.

It can be seen from the figure that the value of partial-contribution value has a weak impact on the final consistency, with a flat curve. Consistency only decreases when partial-contribution value is 0, the same as no contribution, and there is a slight decrease when the value approaches 1 of full contribution. This reflects the importance of the level of partial-contribution.

E.2 Cross Model Sensitivity

Table 8: Agreement between RTE and human trajectory-level judgments under different evaluator configurations. Spearman's ρ remains consistently high across facts extractors and utility annotators, indicating robust human alignment.

Facts Extractor	**Utility Annotator**	**Spearman's ρ**
Qwen2.5-7B-Instruct	GPT-4o	0.7406
Llama3-8B-Instruct	GPT-4o	0.7675
GPT-4o	GPT-4o	0.7804
Qwen2.5-7B-Instruct	Qwen2.5-7B-Instruct	0.7715
Llama3-8B-Instruct	Qwen2.5-7B-Instruct	0.8294
GPT-4o	Qwen2.5-7B-Instruct	0.8361

To assess whether RTE's agreement with human judgments depends on a particular evaluation model, we recompute RTE using different facts extractors and utility annotators. The results are shown in Table 8. Across all configurations, RTE maintains strong rank correlation with human trajectory-level assessments, with Spearman's ρ ranging from 0.7406 to 0.8361. This indicates that RTE's human alignment is robust to reasonable variations in both components of the automatic evaluation pipeline. This robustness stems from our reclassification of the second RTE task into relatively definite and straightforward subtasks: information extraction and semantic analysis.

E.3 Correlation between RTE and Generation Quality

Table 9: Correlation and discriminative statistics between RTE and final correctness (LJ Score) on Original experiment evaluations. For the calculation of AUROC, we perform binary discretizations on the LJ value. If LJ $= 0$, the label is set to 0; if LJ > 0, the label is set to 1.

Backbones	**Pearson r**	**Spearman ρ**	**AUROC**
InternVL3.5-8B	0.2985	0.2646	0.6425
Qwen2.5-VL-7B	0.2089	0.1541	0.6163
Qwen3-VL-4B	0.3478	0.3296	0.7175
Qwen3-VL-8B	0.3713	0.3364	0.7249
Qwen3-VL-32B	0.3370	0.3176	0.7047
GPT-5.1	0.2910	0.3249	0.6834

As shown in Table 9, overall across all models, $r > 0$, $\rho > 0$, and $\mathrm{AUROC} > 0.5$, which indicates that there exists a stable positive correlation between RTE and the correctness of final answers. When the retrieval step can supply more novel evidence that is effectively leveraged by subsequent reasoning, the model is more likely to generate correct answers. Nevertheless, this correlation is not

strong on the whole. This suggests that RTE cannot replace the quality metric of final answers; instead, it offers a complementary process-level perspective: RTE focuses more on measuring whether retrieved evidence truly advances reasoning progress within trajectories, however, the final answers may also be affected by the model's internal knowledge, visual recognition capability and reasoning robustness. Furthermore, the predictive power of RTE does not increase monotonically with model scale. Qwen3-VL-4B and Qwen3-VL-8B achieve high AUROC values, whereas GPT-5.1 scores slightly lower, and the relatively under-performing Qwen2.5-v-7B yields an even lower value.

Overall, these imply that the final answers of certain models may rely more on internal knowledge or visual grounding, reducing the dependence of final correctness on the agent iteration process. **This further corroborates the argument of this paper: the robustness of final answers is not fully consistent with trajectory robustness.**

Table 10: Explaining RTE–LJ misalignment with Direct VLM. We evaluate Direct VLM on ReAct trajectory subsets defined by RTE and final-answer correctness. The high recovery rate on Low-RTE-Correct samples shows that many correct answers with inefficient retrieval trajectories can be attributed to retrieval-independent backbone knowledge, supporting the complementary roles of RTE and LJ.

Subsets	**Mean RTE**	**ReAct Correct**	**Direct VLM Correct**
Low-RTE-Correct	0.000	1.000	0.791
Low-RTE-Wrong	0.004	0.000	0.209
High-RTE-Correct	0.824	1.000	0.492

Furthermore, we conduct an in-depth analysis of the reasons underlying the weak correlation between RTE and LJ. Using Qwen2.5-VL-7B, we construct high and low trajectory efficiency subsets by grouping samples with RTE scores in the top 25% and bottom 25% respectively, and compare the accuracy of answers obtained via the ReAct paradigm against those directly generated by the VLM within these subsets. The results are presented in Table 10. It can be observed that for the subset with low RTE yet high LJ scores, direct VLM inference still yields strong accuracy. Meanwhile, samples with low RTE and low answer accuracy exhibit similarly poor performance when relying solely on direct VLM generation. This experiment sufficiently demonstrates that high-confidence samples for which the VLM inherently holds the correct answer can still be accurately resolved using internal knowledge even with inefficient reasoning trajectories, reflecting greater robustness to trajectory quality.

F Experiment Setup Details

We evaluate a ReAct-style pipeline using Qwen2.5-VL-7B-Instruct as the backbone models for main experiments, and Qwen family models including Qwen3-VL series and InternVL3.5 (Wang et al., 2025b) for cross-performance model validation within the same family. We reproduce ReAct on FlashRAG (Jin et al., 2025b) and re-engineered the prior ReAct implementation from (Li et al., 2024). Especially, we restrict the action space to textual and image retrieval while enhancing trajectory robustness by LLM output pruning. Each iteration proceeds through: action generation, action parsing, conditional retrieval, evidence injection, and subsequent generation. We set the maximum iteration rounds to 5. The full trajectory—including all triggered actions, inputs/outputs, and overall trajectory states—is meticulously logged to facilitate interventions via static query rewriting and context restoration. Trajectory states are categorized as **OK**, **Max-Turn Reached** (MTR), or **Generation Error** (GE) induced by context overflow. In the latter two cases, we prompted the LLM to summarize the preceding trajectory and produce the conclusions already discovered as the final answer.

Table 11: Referential under-specification occurrence frequency across backbone models and task types. Due to the unique training characteristics of MMSearch-R1, the trajectory that triggers retrieval only initiates a single retrieval, resulting in identical frequencies under both the Item and Query scales.

Model	Comp.		E.R.		M.H.		S.A.Q.		C.P.R.	
	Item	Query	Item	Query	Item	Query	Item	Query	Item	Query
InternVL Series										
InternVL3.5-8B	48.32	44.30	59.44	52.75	53.20	49.30	73.65	64.47	47.65	33.99
Qwen2.5-VL-Instruct Series										
Qwen2.5-VL-7B	33.45	29.98	49.70	43.13	41.95	33.35	51.46	47.29	25.81	16.67
Qwen3-VL-Instruct Series										
Qwen3-VL-4B	22.15	15.13	42.63	34.24	22.41	16.93	45.42	37.36	27.82	13.58
Qwen3-VL-8B	22.74	12.88	46.08	32.57	23.59	13.02	49.65	38.65	24.65	13.57
Qwen3-VL-32B	4.71	2.65	24.80	11.62	9.51	3.36	11.16	8.98	1.18	0.50
Qwen3.5 Series										
Qwen3.5-4B	7.79	2.64	26.32	15.64	10.29	4.00	6.33	3.76	3.43	1.60
Qwen3.5-9B	31.08	13.20	40.26	29.08	30.86	13.60	41.10	23.86	15.42	10.48
Qwen3.5-27B	1.20	0.38	11.67	5.06	3.57	2.12	8.97	3.98	4.55	2.37
Reinforcement Learning Series										
MMSearch-R1-7B	8.36	8.36	27.07	27.07	6.82	6.82	3.23	3.23	0.91	0.91
Proprietary Series										
GPT-5.1	6.06	4.28	16.07	13.58	14.50	7.80	14.29	12.11	3.07	2.12

G More Findings

G.1 Entity Ambiguity Frequency across Different VLM Backbones

Table 11 reports the frequency of referential under-specification for additional backbone models categorized by task types and dataset sources. At the item level, an item is considered under-specified if **at least one labeled retrieval query** within its trajectory is entity under-specified. At the query level, the frequency is computed as the ratio of entity under-specified labeled retrieval queries to the total number of labeled retrieval queries across all items. Referential under-specification appears consistently across models of different families and scales, as well as across diverse VQA tasks. InternVL3.5-8B exhibits a high frequency of under-specified references. This indicates that under-specified reference is a common issue across models rather than a unique characteristic of a single model. We also conducted statistics on models of different sizes and found that for the Qwen3 series, the phenomenon of reference under-specification decreases as the model scale increases. However in the Qwen3.5 family: **while the 4B and 27B variants display very low probability of entity under-specification, the 9B variant exhibits substantially higher likelihood.**

G.2 Analysis of Trigger Conditions for Under-specified Reference

Before studying how referential definiteness affects trajectories, we first examine when referential under-specification is most likely to arise in RefAmb. Figure 8 illustrates the occurrence frequency of referential under-specification under different combinations of C_1 (modality coupling) and C_2 (external evidence dependency). Notbly, the combination of $C_1 =$ Yes and $C_2 =$ No does not exist in RefAmb. The results indicate that samples with strong multimodal coupling and high reliance on external knowledge are most prone to referential under-specification. The `N/N` combination also yields a relatively higher frequency of referential under-specification than `Y/Y`, as entity recognition for such samples still depends on visual context. Nevertheless, since their answers generally do not require support from external evidence, the model has little need to invoke textual retrieval tools. Consequently, this type of referential under-specification is rarely exposed at the query–retriever interface and exerts a comparatively limited impact on retrieval trajectories.

G.3 Trajectory Length Distribution

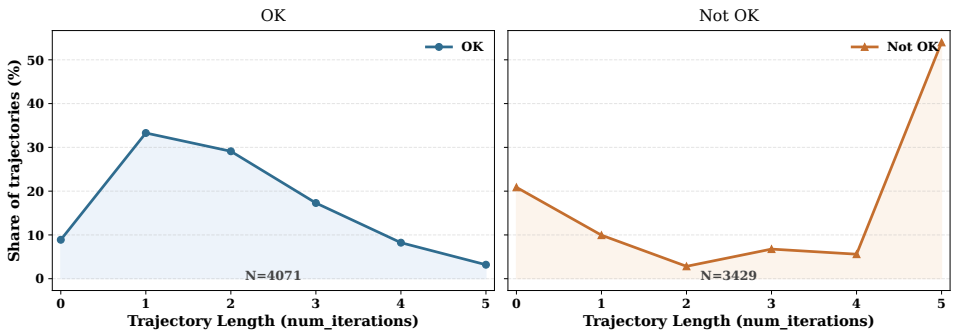

Figure 9: The trajectory length distributions of two states. **Left:** the trajectory length distribution when the status is OK (i.e., the final answer is exported). **Right:** presents the trajectory length distribution when the status is failed and the answer has to be summarized to terminate the process. The curves are aggregated statistics from all Backbone models used in the experiment.

We analyze trajectories based on status, where `status=OK` demonstrates that a trajectory successfully iterates to the predefined answer to generate an action. We separately count the changes in trajectory activity corresponding to the two states `MTR` (Maximum Turns Reached) and `GE` (Generation Error by context overflow) under `status=OK` and `status != OK`. Figure 9 illustrates the overall trajectory length distributions with Qwen2.5-VL-7B-Instruct, Qwen3-VL-4B, 8B, 32B Instruct and InternVL3.5-8B serving as Backbones for both status conditions.

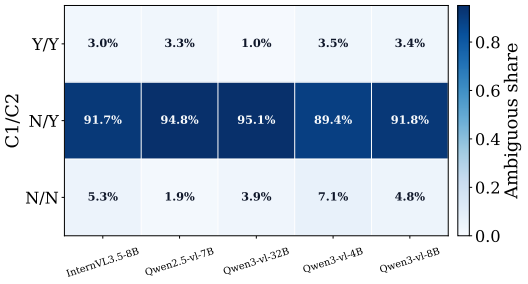

Figure 8: Relationship between dependency tags combination and the frequency of referential under-specification. We calculate the two-dimensional distribution of C_1 and C_2 within the subset of samples with ambiguous references.

It can be observed from Figure 9 that the lengths of trajectories with an OK status that converge successfully to the final answer are mostly concentrated in the range of 1 to 3. This implies that questions correctly answered by the VLM have little reliance on exploratory processes such as retrieval, with answers derived primarily from its internal knowledge. While these trajectories complete execution successfully, they fail to fully leverage the exploratory capabilities of the agent and instead obtain results directly from embedded parametric knowledge. In contrast, trajectories do not marked OK generally feature longer lengths that reach the predefined `Max Iteration`.

G.4 Residual Ambiguity after Oracle Intervention

Table 12: Entity under-specification before and after Oracle intervention.

Metric	Before Oracle intervention	After Oracle intervention	Reduction
Item Scale	100.00%	55.67%	44.33 pct
Query Scale	64.47%	34.79%	29.68 pct

Table 12 shows that residual under-specification remains after first-round oracle intervention. Across paired cohorts of original and Oracle-intervened trajectories, items containing at least one entity under-specified retrieval step decrease from 100% to 55.67%, and under-specified retrieval steps drop from 64.47% to 34.79%. The persistence of residual under-specification demonstrates that first-round Oracle intervention does not merely rectify a one-off input defect. This pattern suggests a mismatch between the VLM and the retriever: the VLM does not fully account for the retriever's interface, preferences, and feedback when generating queries.

G.5 Details on RTE and LJ across Baselines under Two Interventions

Figure 2 illustrates the relative changes in the impacts of the two interventions on trajectories, and Figure 10 presents the absolute values of RTE and LJ under three settings. We utilize these raw scores to analyze the performance of different task categories in terms of RTE and LJ.

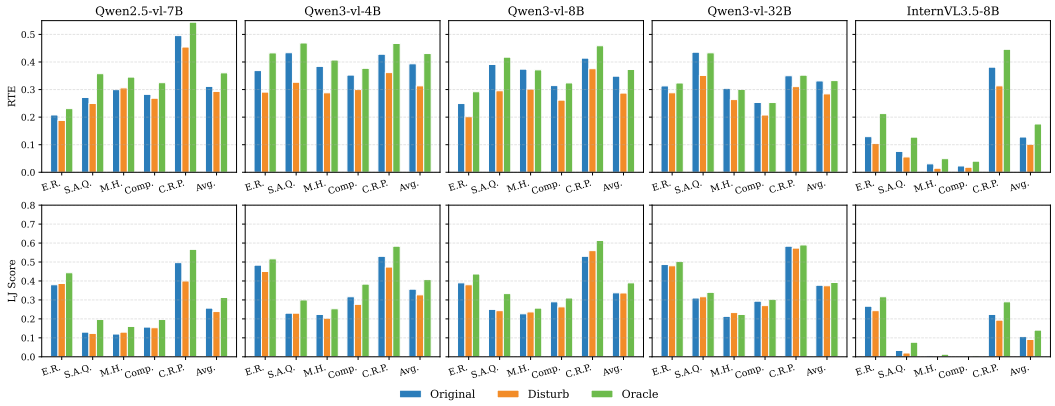

Figure 10: Score variations of RTE and LJ values across different models and task categories under two intervention methods.

The joint distribution of C_1 and C_2 in Figure 6 helps explain the task-wise patterns of RTE and LJ. In E.R., a large fraction of samples fall into $C_1 = \text{No}, C_2 = \text{No}$. These cases rely more on visual perception or parametric knowledge than on retrieval, leading to relatively high LJ but low trajectory activity. In contrast, M.H. has the highest proportion of $C_1 = \text{No}, C_2 = \text{Yes}$ samples, which corresponds to lower LJ scores. Comp. and C.R.P. contain more $C_1 = \text{Yes}, C_2 = \text{Yes}$ cases, indicating stronger textual anchoring. However, C.R.P. exhibits higher trajectory efficiency and generation quality than Comp., This indicates difficulty of Comp. shows stronger robustness to weak cross-modal coupling, since comparative questions are often organized around multiple explicit entities, whereas C.R.P. remains one core entity in most cases.

H Algorithms of IIGCR

Algorithm 1: Rebuild Messages Module

Input: Messages $\mathcal{M}$, mode $m \in \{\texttt{local_repair}, \texttt{global_reflection}\}$, reference query q, feedback f, Rewrite Prompt $\mathcal{P}$, global_reflection Prompt $\mathcal{Q}$
Output: Rebuilt messages $\mathcal{M}'$, rebuilt response a

```
// mode = local_repair
```
if $m = \texttt{local_repair}$ **then**
  $a_{raw} \leftarrow \text{LLM}(q, f, \mathcal{P})$;
  $(d, a) \leftarrow \text{ValidationOrRectification}(q, a_{raw}, f)$;
  $\mathcal{M}' \leftarrow \mathcal{M}$;
  if *latest* `assistant <Search>` *node exists in* $\mathcal{M}$ **then**
    replace that node with a;
  else
    append node a to $\mathcal{M}'$;
  return $\mathcal{M}', a$;
```
// mode = global_reflection
```
else
  $\widetilde{\mathcal{M}} \leftarrow \text{ContextPruning}(\mathcal{M})$;
  $\mathcal{M}' \leftarrow \widetilde{\mathcal{M}} \oplus \mathcal{Q}$;
  $a \leftarrow \text{LLM}(\mathcal{M}')$;
  return $\mathcal{M}', a$;

As discussed before, IIGCR is a plug-and-play approach that embeds an information-increment-based contextualizer into the agent loop to guide VLMs to correct errors, reflect and replan.

Algorithm 1 shows the logic of Rebuild Messages module, operationalizes these interventions in two distinct modes. In the `local repair` mode, the module employs large language model (This paper uses models with the same VLM Backbone) to generate candidate queries or responses aiming at recovering missing information, followed by a validation or rectification step to ensure semantic consistency and information diversity. The updated message sequence is then incorporated into the trajectory, either replacing the latest retrieval node or appending a new assistant node. In the `global reflection` mode, a stronger corrective procedure is enacted, which prunes recent low-gain retrieval turns and directs the model toward alternative subproblems or clues, effectively enabling a controlled rollback and reorientation of the reasoning trajectory.

Algorithm 2: Validation Or Rectification Module

Input: Reference query q, raw rewritten response a_{raw}, feedback f
Output: Rectified response a
extract $\hat{q}$ from the <Search> block of a_{raw};
if $\text{sim}(\hat{q}, q) \leq \tau_s$ **then**
  $a \leftarrow a_{\text{raw}}$;
else
  if $\text{AmbiguityCheck}(q) =$ *True* **then**
    $a \leftarrow$ <Search> Image Retrieval: q </Search>;
return a;

Algorithm 2 illustrates the validation and rectification module. This module computes the similarity between the original query and the rewritten query. If the similarity score exceeds the τ_s threshold, the query rectification is deemed unsuccessful. We then first detect whether reference ambiguity exists in the input question. If reference ambiguity is confirmed, it proves that such ambiguity degrades the retrieval trajectory efficiency of the task. Accordingly, the retrieval mode is switched to image retrieval, enabling the VLM to perform specification on entity references.

Algorithm 3: Stack-based Context Pruning

Input: Messages $\mathcal{M}$, duplicate query q^{dup}, zero-gain query q^0, repaired query $\hat{q}$
Output: Pruned messages $\widetilde{\mathcal{M}}$
Initialization: context stack $\mathcal{S} \leftarrow [\,]$, removed stack $\mathcal{D} \leftarrow [\,]$, index $i \leftarrow 1$;
if $q^{dup} = \varnothing$ ***and*** $q^0 = \varnothing$ **then**
  return $\mathcal{M}$;
while $i \leq |\mathcal{M}|$ **do**
  $u \leftarrow \mathcal{M}[i]$; $\text{Push}(\mathcal{S}, u)$;
  if $\text{Top}(\mathcal{S})$ *is* `assistant` ***and*** *text-retrieval* ***in*** *<Search>* **then**
    $q_u \leftarrow \text{ExtractQuery}(\text{Top}(\mathcal{S}))$;
    if $q_u = q^{dup}$ ***or*** $q_u = q^0$ **then**
      $u^{bad} \leftarrow \text{Pop}(\mathcal{S})$;
      $\text{Push}(\mathcal{D}, u^{bad})$;
      `// Discard retrieval feedback following the invalid search`
      if $i+1 \leq |\mathcal{M}|$ ***and*** $\mathcal{M}[i+1]$ *is retrieval-feedback* `user` *message* **then**
        $\text{Push}(\mathcal{D}, \mathcal{M}[i+1])$;
        $i \leftarrow i+1$;
      `// Discard the induced thought conditioned on invalid feedback`
      if $i+1 \leq |\mathcal{M}|$ ***and*** $\mathcal{M}[i+1]$ *is* `assistant` *<Thought>* **then**
        $\text{Push}(\mathcal{D}, \mathcal{M}[i+1])$;
        $i \leftarrow i+1$;
      `// Replace the invalid retrieval state with a repaired query`
      $\hat{u} \leftarrow \text{BuildSearchMessage}(\hat{q})$;
      $\text{Push}(\mathcal{S}, \hat{u})$;
  $i \leftarrow i+1$;
$\widetilde{\mathcal{M}} \leftarrow \text{List}(\mathcal{S})$; **return** $\widetilde{\mathcal{M}}$;

Algorithm 3 shows the Context Pruning module, This module models the context trajectory sequence of the VLM as a stack structure for maintaining message hygiene by selectively filtering out redundant or uninformative elements. It iterates through the message history, identifies assistant retrieval nodes corresponding to repeated queries or zero-information steps, and pops them from the active trajectory stack, preserving only essential and novel contributions. This module ensures that the trajectory remains tractable and that subsequent retrievals are evaluated against a concise, high-quality context, thereby preventing cumulative noise from impeding reasoning progress.

Collectively, these three components establish a coherent framework for monitoring, repairing, and pruning agentic retrieval trajectories, balancing exploration with structural integrity, and mitigating the propagation of low-value or redundant retrievals while maximizing the effective utilization of newly acquired evidence. The prompts involved in the above algorithm can be found in Appendix L.8.

I Additional Analysis of IIGCR

I.1 Hyper-Parameter Sensitivity

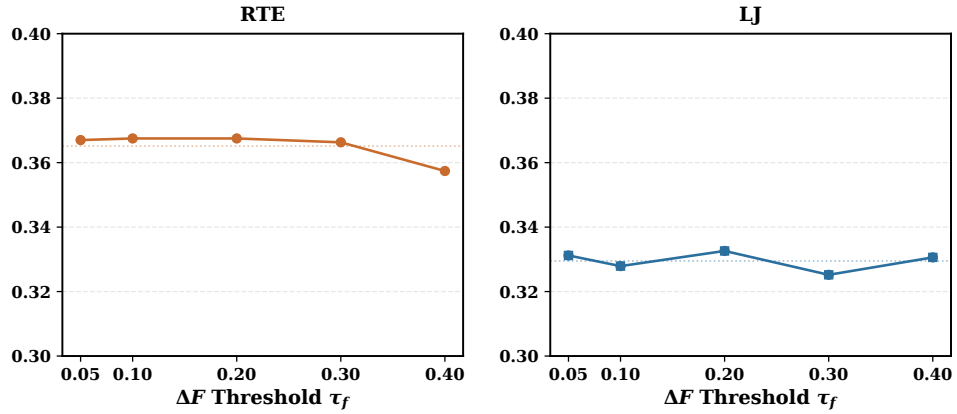

Figure 11: Sensitivity analysis of the early warning threshold τ_f with low ΔF, conducted with $\tau_s = 0.90$.

We conduct a sensitivity analysis on the two most critical core parameters required in IIGCR: the ΔF threshold τ_f and the Similarity threshold τ_s.

Low ΔF Threshold. As shown in Figure 11, overall, the IIGCR method exhibits strong robustness to the value of τ_f: the RTE value remains stable around 0.36, with only a slight drop occurring at 0.40; LJ stays within the range of 0.32 to 0.33 and shows minor fluctuations. This suggests that IIGCR operates in a regime where low τ_f detection is structurally stable, reducing sensitivity to heuristic threshold design.

Query Similarity Threshold. In contrast to τ_f, the system shows a mild but more noticeable sensitivity to the query similarity threshold τ_s. As shown in Figure 12, the overall performance exhibits a slight upward trend as τ_s increases. This behavior suggests that when τ_s is low, the system may over-trigger context reconstruction even for moderately similar queries, which can introduce unnecessary adjustments. In contrast, a higher τ_s effectively filters out such cases and focuses IIGCR on genuinely repetitive or stagnant trajectories, leading to more stable improvements.

As shown in Table 3, the IIGCR intervention delivers consistent improvements in trajectory success rates for both the Qwen2.5 and Qwen3 models, while markedly reducing the average number of iteration steps. This demonstrates its ability to effectively alleviate the stagnation phenomenon in retrieval trajectories. Furthermore, we conducted a state transition analysis on the subset of samples where both the Oracle and IIGCR intervention methods were activated simultaneously. As illustrated in Figure 13, the transition distribution of IIGCR exhibits a pronounced OK-state bias: compared

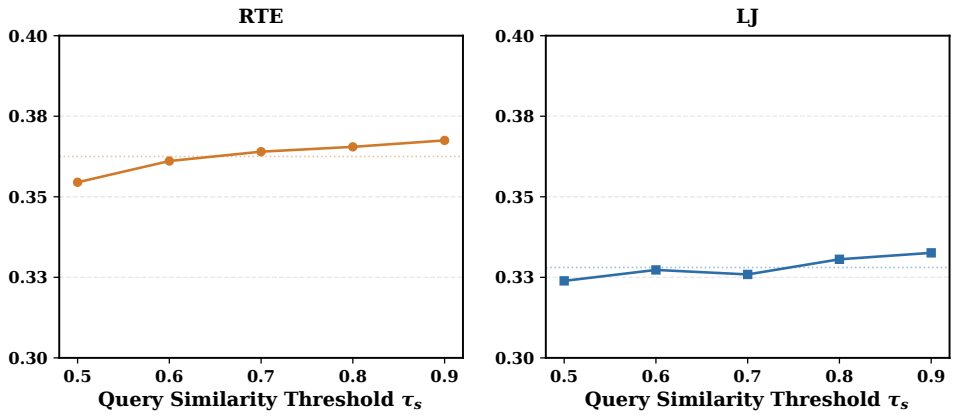

Figure 12: Sensitivity analysis of the early warning threshold τ_s with high query similarity, conducted with $\tau_f = 0.20$.

with Oracle, IIGCR enables a larger proportion of originally failed trajectories to transition into successful ones. This improvement is especially prominent on Qwen3, indicating that IIGCR can not only fix local retrieval failures but also more effectively break trajectory stagnation caused by repeated queries or low information gain, thereby exerting a steady positive effect on mitigating stagnation.

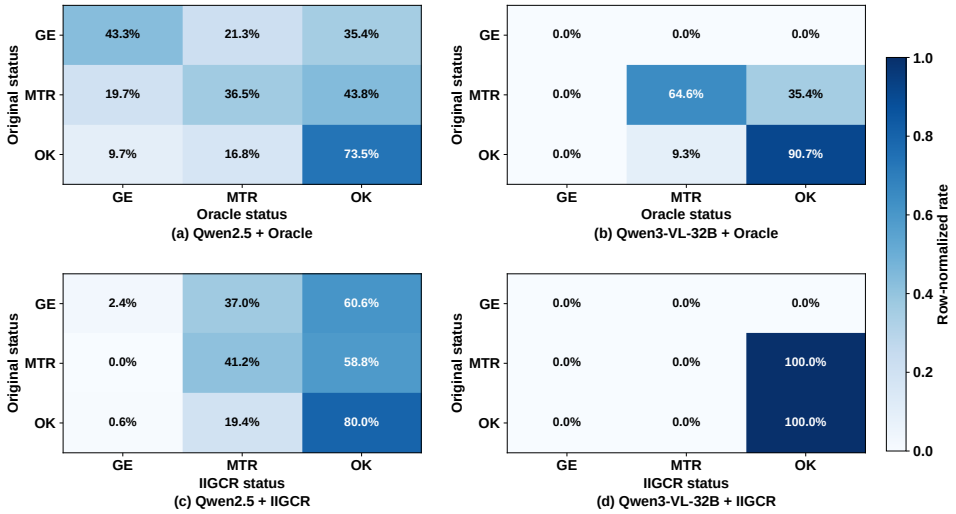

Figure 13: Comparison of the intervention effects of the Transition Matrix of IIGCR on two types of Backbones with the Oracle. We conducted the analysis on the subset of samples where both the Oracle and IIGCR intervention methods were activated simultaneously.

I.2 Ablation Study

We conduct ablation studies on the Rebuild Messages module and ROI-guided Image Retrieval, with results reported in Table 13. The results demonstrate that the components of IIGCR jointly contribute to improvements in both trajectory efficiency and generation quality. Specifically, Global Reflection is the key component for maintaining generation quality, while Context Pruning primarily improves reasoning efficiency by removing redundant context with minimal impact on generation quality. This reveals that context pruning does not impair VLM task execution, while reduces the overhead of context window expansion.

Table 13: Ablation on IIGCR with Qwen2.5-VL-7B-Instruct as backbone. The IIGCR is in $\tau_s = 0.90$ and $\tau_f = 0.20$ hyper-parameter setting.

Method	E.R.		S.A.Q.		M.H.		Comp.		C.R.P		Avg.	
	RTE	LJ	RTE	LJ	RTE	LJ	RTE	LJ	RTE	LJ	RTE	LJ
IIGCR	20.89	48.33	35.09	25.00	37.02	16.33	35.24	28.00	51.99	50.00	36.05	33.53
w/o Local Repair	20.89	46.67	34.18	23.33	35.79	16.33	34.40	27.33	51.60	50.00	35.37	32.73
w/o Global Reflection	18.83	44.67	34.95	21.67	36.82	15.67	35.85	26.33	52.66	48.33	35.82	31.33
w/o Context Pruning	18.44	46.33	31.90	24.00	35.74	16.33	35.38	27.00	51.60	50.33	34.61	32.80
w/o All Above	18.32	42.33	34.54	21.67	34.82	14.00	35.34	23.33	50.63	49.00	34.73	30.07
w/o ROI	20.58	48.00	35.75	23.67	35.02	16.00	35.26	27.33	52.32	48.67	35.79	32.73

Local Repair acts as a stabilization mechanism that corrects suboptimal intermediate trajectories, where its removal leads to moderate degradation across both metrics. ROI-guided retrieval provides a uniform and mild enhancement, contributing consistently across settings without introducing task-specific bias. Furthermore, we observe **clear synergistic effects** among Local Repair, Global Reflection, and Context Pruning. Removing all three components is equal to eliminate the context reconstruction mechanism in IIGCR, resulting in consistent degradation in both trajectory efficiency and generation quality. This suggest that the modular design of IIGCR reduces sensitivity to individual components, preventing catastrophic performance collapse under single-module ablation. Instead, the interaction among modules leads to a complementary effect, yielding performance gains that are greater than the sum of individual contributions.

J Theoretical Analysis on RTE

Proposition 1. $\mathrm{RTE} \in [0, 1]$. *And the overall RTE will approach 1 only when sufficient and sufficiently useful new information is retrieved in each iteration round.*

Proof. We first analyze the range of S_t. According to equation 1,

$$\Delta F = \frac{|F_t \cup F_{<t}|}{|F_t| + \epsilon} - \frac{|F_{<t}|}{|F_t| + \epsilon} \leq \frac{|F_t| + |F_{<t}|}{|F_t| + \epsilon} - \frac{|F_{<t}|}{|F_t| + \epsilon}. \tag{5}$$

Besides,

$$|F_t \cup F_{<t}| - |F_{<t}| \geq 0, \tag{6}$$

that is,

$$0 \leq \Delta F \leq \frac{|F_t|}{|F_t| + \epsilon} \leq 1. \tag{7}$$

Also,

$$0 \leq U_t \leq 1. \tag{8}$$

Thus, $0 \leq S_t \leq 1$, hence $0 \leq \mathrm{RTE} \leq 1$. And $\mathrm{RTE} = 1$ iff

$$\forall \tau \in \mathcal{T}, U(\tau) = 1 \wedge F_\tau \cap F_{\tau-1} = \varnothing, \tag{9}$$

$\mathrm{RTE} = 0$ iff

$$\forall \tau \in \mathcal{T}, U(\tau) = 0 \vee F_\tau \subset F_{\tau-1}. \tag{10}$$

□

K Case Study

Table 14 shows a representative failure-recovery example from Qwen2.5-VL-7B on RefAmb. The question asks for the weight of the male bird in a blurry water-bird image, where the gold answer is 1.729 kg. This sample is useful because the visual evidence is not strong enough to make the species obvious, so the trajectory must first identify the bird and then retrieve the correct attribute. It

Table 14: Case study comparing Original, Oracle, and IIGCR on a Qwen2.5-VL-7B trajectory. The sample asks for the weight of the male bird shown in the image. Oracle introduces the correct entity hint but still stalls on repeated generic identification queries, whereas IIGCR repairs the trajectory into a targeted attribute lookup after stagnation is detected.

Image	**Question / Gold answer**	
	**Question**	What is the weight of a male of this bird in kilogram?
	Gold answer	1.729 ([1.5561, 1.9019]) kg (male Red-throated loon).

Setting	**Trajectory Record (Concise)**	**Outcome**
Original	<Search> Identify the bird species in the image. </Search> <Search> Weight of a male Least Grebe in kilograms. </Search> <Search> Weight of a male Least Grebe in kilograms. </Search> **<Voice-over> The trajectory repeatedly searches the same incorrect attribute query and fails to produce a final answer. </Voice-over>**	missing_final_answer
Oracle	<Search> Red-throated loon. </Search> <Search> bird in the image. </Search> <Search> bird in the image. </Search> **<Voice-Over> Although the oracle entity is injected, the trajectory does not preserve this anchor and falls back to a generic query loop. </Voice-Over>**	missing_final_answer
IIGCR	<Search> Identify the species of the bird in the image. </Search> <Search> Identify the species of the bird in the image. </Search> <Controller> similar_query_streak $\geq$ 1 and judgement not passed. </Controller> <Action> ROI + Image Retrieval repair. </Action> <Search> Average weight of a male Red-throated loon in kilograms. </Search> <Search> Average weight of a male Red-throated loon in kilograms. </Search> <Controller> similar_query_streak $\geq$ 1. </Controller> <Search> Characteristics of a male Red-throated loon. </Search> <Final Answer> Approximately 1.5–2.0 kg. </Final Answer>	OK

therefore exposes whether the controller can redirect a stalled search after the initial grounding step still fails to produce an answer.

Under the **Original** setting, the model first drifts to *Least Grebe* and then repeatedly asks for *"Weight of a male Least Grebe in kilograms"*, even though that branch does not answer the question. **Oracle** replaces the first-step under-specified referent with *Red-throated loon*, but the trajectory still stagnates: after an incorrect detour through *"Toxostoma"*, it falls back to the generic query *"bird in the image"* and repeats it for the remaining rounds. This example highlights the limitation of static anchoring alone: even when Oracle supplies the right entity hint, it does not repair the trajectory once the search collapses into a low-information loop.

By contrast, **IIGCR** explicitly repairs the stalled trajectory. It first reaches the species-level clue *Red-throated loon*, then moves to the attribute query *"Weight of a male Red-throated loon in kilograms."*. When a `similar_query_streak>=1` pattern appears, the controller rewrites that request into the more direct form *"Characteristics of a male Red-throated loon."*. The top-1 retrieval result of the revised query is an encyclopedia entry containing information about the weight of the Red-throated loon, enabling the model to answer this question. Finally, the model outputs an approximate answer of *1.5–2.0 kg*, which is consistent with the gold value 1.729 kg. Relative to Oracle, the key gain is not just better grounding, but the ability to convert a repeating species-identification loop into a targeted attribute lookup once stagnation is detected.

L Prompt Templates in the Whole Research Project

We have provided all the important prompts involved in completing this research project. Due to space limitations, only the core instructions of the prompts are presented. To keep the paper concise, only the key few-shot examples or constraints in the prompts are retained.

L.1 Prompt for Slot Filling in RefAmb Curating

Prompt: Task Type Categorization Prompt

Input Prompt:

- Text Query: {`text_query`}

You are annotating image-question pairs from InfoSeek for entity-ambiguity analysis in multimodal retrieval-augmented question answering.
Your task is to assign exactly one label to the current sample.
You must judge based on both: 1. the Text Query 2. the Image Query
Do not answer the question itself. Do not use outside knowledge.
Label set
Choose exactly one: - Entity Recognition - Single-hop Attribute Query - Others
Definitions
1. Entity Recognition Use this label when the main goal is to identify who or what the depicted entity is at a sufficiently specific level that uniquely identifies the intended entity.
This includes cases where the answer is primarily: - the specific name of the object, building, species, plant, vehicle model, place, artwork (its name or title), or brand - the specific identity of the depicted entity itself
Use Entity Recognition only when the core question is "what exactly is this?" or "who/which entity is this?"
Important: - Entity Recognition requires identity at a level specific enough to distinguish the target entity, not just a broad visual class. - For example, identifying something only as "a sedan", "a church", "a bridge", or "a flower" is usually too coarse if the intended task is to identify which sedan, which church, which bridge, or which flower it is. - If the question asks only for a generic category, type, family, or broad visual description, do not use Entity Recognition unless that generic type is clearly the intended final identity level. - If the question asks for one fact about an already-grounded entity, do not use Entity Recognition. - Treat the following as identity-defining fields that should still be labeled Entity Recognition when they are the main target of the question: brand, model, proper name, title, species, breed, exact

building name, exact place name. - Entity Recognition is only for the identity of the entity itself, not for one fact about that entity.
2. Single-hop Attribute Query Use this label when the image helps ground the entity, but the actual question asks for one direct fact, attribute, or relation about that entity.
Typical answer types include: - style - manufacturer - country - mountain range - taxonomy - date - size - material - location - category - affiliation - dedication target - endemic region - protected area - parent group or related organization
Important: - Many InfoSeek questions mention "this building", "this mountain", "this vehicle", or "this plant". - If the question is not asking for the identity of the entity itself, but for one direct fact about it, label it as Single-hop Attribute Query. - Questions asking for one directly related entity (such as manufacturer, country, parent group, mountain range, affiliation, etc.) should also be labeled Single-hop Attribute Query in this coarse taxonomy. - Questions asking where the entity is located, what region/country it belongs to, what it is dedicated to, what protected area it is in, where it is endemic to, or another single direct fact about it should be labeled Single-hop Attribute Query. - Do not use Single-hop Attribute Query for identity-defining targets such as brand, model, proper name, title, species, or breed when the main goal is to identify the entity itself.
3. Others Use this label when the main goal is neither Entity Recognition nor Single-hop Attribute Query.
Important rules
- Focus on the main task required to answer the question. - Do not classify based only on wording; use the image to judge whether the question is asking for the entity identity itself or for a fact about the entity. - For Entity Recognition, prefer the finest identity level needed to uniquely identify the depicted entity rather than a broad category-level description. - If the question asks for identity-defining information such as brand, model, proper name, title, species, or breed, prefer Entity Recognition. - If the question asks for an attribute, property, affiliation, category, relation target, location, style, origin, or manufacturer of the depicted entity, prefer Single-hop Attribute Query. - If the question asks for dedication target, endemic region, protected area, region, country, affiliation, location, or another single relation/fact about the depicted entity, prefer Single-hop Attribute Query. - If the question asks what the depicted entity itself is, prefer Entity Recognition. - Return only valid JSON.
Output format:
`{"coarse_label": "Yes" or "No", "rationale": ""}`
Where: - coarse_label must be exactly one of: - Entity Recognition - Single-hop Attribute Query - Others - rationale should be one concise sentence - confidence should be an integer from 0 to 100
Example 1 Question: What is this plant? Expected JSON: "coarse_label": "Entity Recognition", "rationale": "The main goal is to identify which specific plant the depicted entity is rather than ask for a fact about it."
Example 2 Question: What is the architectural style of this building? Expected JSON: "coarse_label": "Single-hop Attribute Query", "rationale": "The image grounds the building, but the question asks for one direct attribute of it."

L.2 Prompts for Referential Dependency Annotation

Prompt: C_1 Text-Alone Entity Identification Prompt

Input Prompt:

- Question: {`question`}

Task: Decide whether the core target entity can be reliably identified from the question text alone.
Label set:

- `Yes`
- `Partially`
- `No`

Decision rule:

- Use `Yes` if the question text explicitly names the entity or gives enough information to reliably identify it.
- Use `Partially` if the text gives some descriptive clues, but not enough to uniquely identify the entity for retrieval.
- Use `No` if the text alone does not provide usable information to identify the entity.

Output format:

```
{"focused_entity": "...", "text_alone_identifies_entity": "...",
↪  "rationale_1": "..."}
```

Prompt: C_2 Stage-1 Visual Evidence Extraction Prompt

Input Prompt:

- Question: {`question`}

Task: Extract only the task-relevant evidence that is directly observable in the image. Do not answer the question.

Output format:

```
{"relevant_visible_evidence": [],
↪  "uncertain_or_missing_visual_points": [], "extractor_summary": ""}
```

Prompt: C_2 Stage-2 External Evidence Dependency Prompt

Input Prompt:

- Question: {`question`}
- Gold Final Answer: {`gold_answer`}
- Stage-1 Visual Evidence: {`evidence`}

Task: Judge whether the final answer is uniquely recoverable from the image evidence alone.

Label set:

- `Yes`
- `No`
- `Uncertain`

Decision rule:

- Use `Yes` if the image evidence alone is sufficient to recover the final answer.
- Use `No` if the evidence is compatible with the answer but not sufficient to uniquely determine it, or if the answer requires evidence beyond the image.
- Use `Uncertain` if the evidence summary is incomplete or under-specified enough that a confident decision cannot be made.

Output format:

```
{"answer_recoverable_from_image_alone": "...",
↪  "external_evidence_dependency": "...", "missing_information_type":
↪  [], "rationale": "..."}
```

L.3 Prompt for Ambiguous Reference Recognition

Prompt: Entity Ambiguity Annotation Prompt

Input Prompt:

- Text Query: {`text_query`}

Task: Determine whether the query contains anaphoric pronouns that can only be understood with reference to images at the query-retriever interface. The retriever only sees the text query, not the image or hidden context.
Decision rule: Mark `"entity_underspecified": "Yes"` if the query does not uniquely specify the intended entity for retrieval. Mark `"entity_underspecified": "No"` if the query already names a concrete retrievable target. Do not confuse answer difficulty with entity under-specification.
Output format:
`{"entity_underspecified": "Yes" or "No", "remark": ""}`

L.4 Prompt for Iteration Utility Annotation

Prompt: Iteration Utility Annotation Prompt

Input Prompt:

- Sub-question: {`sub_question`}
- Next Reasoning Text: {`reasoning`}

Task: Understand the semantics of the `reasoning` text and judge the model's satisfaction with the retrieved results. Return exactly one utility label for the current iteration.
Allowed labels:

- `no_contribution`
- `partial_contribution`
- `full_contribution`

Decision rule:

- Use `no_contribution` if the reasoning expresses the evidence is irrelevant, uninformative, insufficient, or discarded.
- Use `partial_contribution` if the reasoning expresses like "evidence helps only part of the sub-question, narrows the search, or resolves one component, but...".
- Use `full_contribution` if the reasoning conveys the meaning of complete resolution.
- Judge only from the model's own next reasoning text, not from your independent opinion.

Output format:
`{"label": "no_contribution", "reason": ""}`

L.5 Prompt for LLM as Judge

Prompt: LLM as Judge Prompt

Input Prompt:

- Reference Answers: {`golden_answers`}
- Candidate's Answer: {`pred`}

Task: Compare the candidate answer against the reference answers and judge whether it is semantically correct.
Decision rule:

- Output `yes` if the candidate answer matches the reference meaning, even if wording differs.
- Output `no` if the answer is wrong, unsupported, incomplete, irrelevant, empty, or does not directly answer the question.
- Output exactly one token: `yes` or `no`.

Output format:
`yes / no`

L.6 Prompt for VLM Executing ReAct

Prompt: System VLM Prompt

Input Prompt:

- Input Question: {`question`}

Task: Decompose the question into sub-questions and solve them step by step. Use `Search` for retrieval and `Final Answer` only when the response is ready.
Search actions:

- `Image Retrieval`: use an anchored phrase for the entity in the image.
- `Text Retrieval`: use a specific query for missing context.
- `No Retrieval`: if retrieval is unnecessary.

Output format:
Use the XML structure with `<Thought>`, `<Sub-Question>`, `<Search>`, and `<End>`.

L.7 Prompts for Two Interventions

Prompt: Oracle Intervention Prompt

Input Prompt:

- Original Question: {`question`}
- Original Text Query: {`text_query`}
- Ambiguity Level: {`under-specification_level`}
- Answers: {`answers`}

Task: Rewrite an entity under-specified retrieval query into an entity-explicit query grounded in the provided answers. Preserve the original retrieval intent and keep edits minimal whenever possible.
Decision rule:

- For object identification, output the canonical entity name only.
- For description or indirect under-specification, replace the vague referent with the grounded entity name.
- For no-object-involved cases, keep translation queries unchanged when appropriate, otherwise disambiguate the target phrase using the answers.
- If the query is already explicit enough for text retrieval, return it unchanged.

Output format:
`{"rewritten_query": ""}`

Prompt: Disturbance Intervention Prompt

Input Prompt:

- Query: {`query`}

Task: Rewrite the query into a more referentially under-specified version for a controlled perturbation experiment. Preserve the original intent, answer target, and task type, but reduce referential clarity.
Decision rule:

- Replace explicit entity names with vague referential expressions when possible.
- Keep the query close in length and structure to the original.
- Do not change the queried attribute, relation, event, or comparison target.
- If the query is already vague, keep it unchanged rather than making it broader.

Output format:
Return only the rewritten query.

L.8 PROMPT FOR IIGCR

Prompt: Rewrite Similar Text Query Prompt (`local_repair`)

Input Prompt:

- Query To Rewrite: {`query_to_rewrite`}
- Feedback: {`vlm_feedback`}

Task: Recover from a blocked retrieval step by choosing a genuinely new retrieval action. Do not repeat the previous query or produce a near-duplicate paraphrase.
Decision rule:

- Use `Text Retrieval` only if a materially different query is justified.
- Use `Image Retrieval` if the missing clue must be grounded in the image.
- Use `No Retrieval` if retrieval is no longer useful.

Output format:
Use exactly one `<Search>` block.

Prompt: Reflection Prompt (`global_reflection`)

Injected Prompt:

`<ReflectionInstruction>`
Repeated or low-gain text retrieval was detected. The recent repeated retrieval turns have been discarded. The current query path is not working and the current retrieval results are not sufficient to move forward. Failed query: reference_query.
Previous thought before the failed retrieval: failed_thought or '[empty]'
Return exactly one `<Thought>` block first, then return exactly one `<Search>` block or one `<Final Answer>` block. Do not omit the required tags. Do not output more than one action block after the thought block.
You must not ask the same question again, and you must not issue the same or near-duplicate Text Retrieval query as: reference_query.
Choose a genuinely different next step. If you choose Text Retrieval, ask a different question that targets a new missing clue, entity, relation, location, time, or attribute instead of rephrasing the same request.
If a different text query is not well justified, choose "Image Retrieval: `anchored phrase`", "No Retrieval", or give a "Final Answer:" if the evidence is already sufficient.

Do not repeat the same failed plan.
</ReflectionInstruction>

Prompt: Ambiguity Check Prompt

System Prompt:

You are an under-specification detector. Return strict JSON as required.

User Prompt:

You are an annotation assistant for multimodal retrieval planning.
Your job is to decide whether a text retrieval query is entity under-specified for a text retriever.
You will receive only a Text Query. You cannot see the image or use hidden context.

Core criterion:
Mark {"entity_under_specified": "Yes"} if the query does not yet contain a canonical entity name or a sufficiently unique grounded target for direct text retrieval.
Mark {"entity_under_specified": "No"} only if the query is already ready for direct text retrieval without seeing the image.

Return exactly one JSON object:
{"entity_under_specified": "Yes" or "No", "remark": ""}

Text Query: {text_query}

M LIMITATIONS & FUTURE WORK

In this study, we use LLM-as-Judge to measure the final generation quality. Although this evaluation method can place more emphasis on semantic understanding, it may also be affected by model bias. Future work can present a more comprehensive analysis of metrics such as Exact Match and Recall. Besides, this paper proposes intervention methods from the perspective of agent context management, a promising direction is joint optimization of VLMs and retrievers, where reward signals such as referential definiteness and trajectory-level information gain guide more robust query generation.